9. Unsupervised Techniques III: Topic Models

DS-GA 1015, Text as Data
Arthur Spirling

April 6, 2021

Goal

Goal

Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents.

Goal

> *Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.*
>
> Blei, 2012

Goal

> *Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.*
>
> Blei, 2012

Note that in social science we often use the outputs from topic models as a measurement strategy:

Goal

> *Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.*
>
> Blei, 2012

Note that in social science we often use the outputs from topic models as a measurement strategy:

"who pays more attention to education policy, conservatives or liberals?"

Recall: Clustering

Recall: Clustering

Document 1 → Cluster K

Document 2

Document 3

$\vdots$

Document N

Cluster 1

Cluster 2

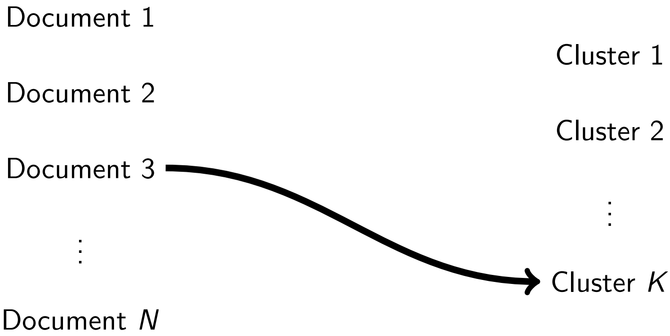
$\vdots$

Cluster K

Recall: Clustering

Document 1	
Document 2	Cluster 1
Document 3	Cluster 2
$\vdots$	$\vdots$
Document N	Cluster K

Recall: Clustering



Recall: Clustering

Document 1

Document 2

Document 3

$\vdots$

Document N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Recall: Clustering

Document 1

Document 2

Document 3

⋮

Document N

Cluster 1

Cluster 2

⋮

Cluster K

Topic Modeling

Topic Modeling

Document 1

Document 2

Document 3

$\vdots$

Document N

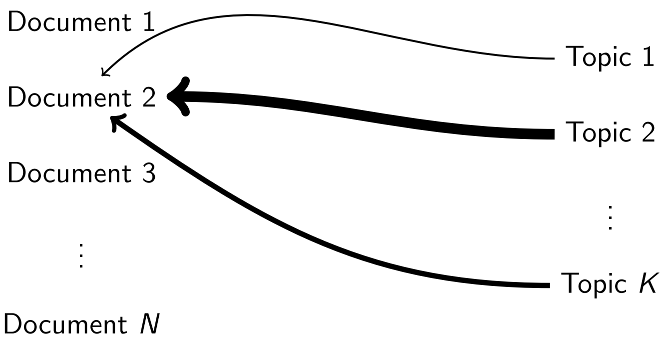
Topic 1

Topic 2

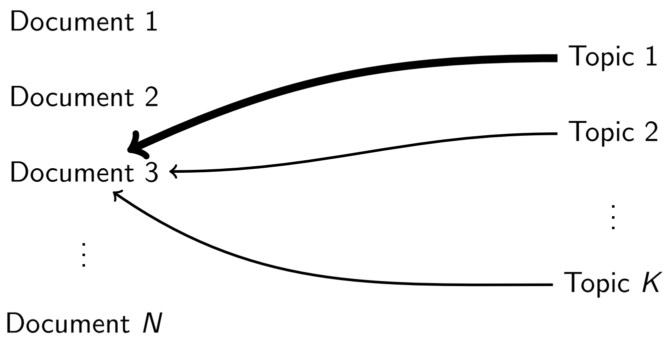
$\vdots$

Topic K

Topic Modeling



Topic Modeling



DGP: intuition

DGP: intuition

Documents exhibit different topics,

DGP: intuition

Documents exhibit different topics, and in different proportions.

DGP: intuition

Documents exhibit different topics, and in different proportions.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

DGP: intuition

Documents exhibit different topics, and in different proportions.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a topic as a distribution over a fixed vocabulary.

DGP: intuition

Documents exhibit different topics, and in different proportions.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a topic as a distribution over a fixed vocabulary.

E.g. the trade topic will have words like `import` and `tariff` with high probability.

DGP: intuition

Documents exhibit different topics, and in different proportions.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a topic as a distribution over a fixed vocabulary.

E.g. the trade topic will have words like `import` and `tariff` with high probability.

Technically we assume the topics are generated first,

DGP: intuition

Documents exhibit different topics, and in different proportions.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a topic as a distribution over a fixed vocabulary.

E.g. the trade topic will have words like `import` and `tariff` with high probability.

Technically we assume the topics are generated first, and the documents are generated second (from those topics).

DGP: intuition

Documents exhibit different topics, and in different proportions.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a topic as a distribution over a fixed vocabulary.

E.g. the `trade` topic will have words like `import` and `tariff` with high probability.

Technically we assume the topics are generated first, and the documents are generated second (from those topics).

Now, where do the words in the documents come from?

Intuition: Generating Words

Intuition: Generating Words

For each document. . .

Intuition: Generating Words

For each document...

1. Randomly choose a distribution over topics.

Intuition: Generating Words

For each document...

1. Randomly choose a distribution over topics. That is, choose one of many multinomial distributions, each which mixes the topics in different proportions.

Intuition: Generating Words

For each document...

1. Randomly choose a distribution over topics. That is, choose one of many multinomial distributions, each which mixes the topics in different proportions.

2. Then, for every word in the document...

Intuition: Generating Words

For each document...

1. Randomly choose a distribution over topics. That is, choose one of many multinomial distributions, each which mixes the topics in different proportions.
2. Then, for every word in the document...
    1. Randomly choose a topic from the distribution over topics from step 1.

Intuition: Generating Words

For each document...

1. Randomly choose a distribution over topics. That is, choose one of many multinomial distributions, each which mixes the topics in different proportions.

2. Then, for every word in the document...
   1. Randomly choose a topic from the distribution over topics from step 1.
   2. Randomly choose a word from the distribution over the vocabulary that the topic implies.

First Part

First Part

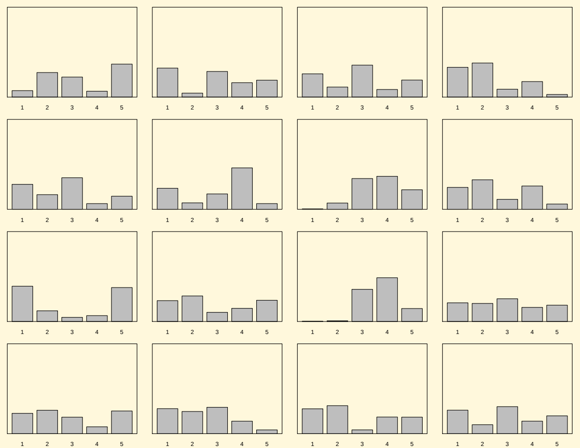
Randomly choose a distribution over topics.

First Part

Randomly choose a distribution over topics. That is, choose one of many multinomial distributions, each which mixes the topics in different proportions.

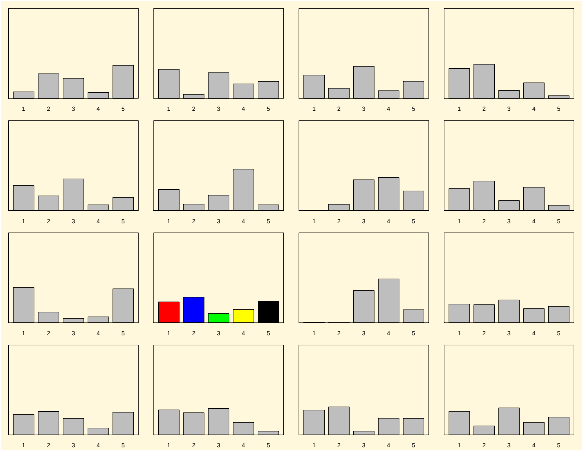
First Part

Randomly choose a distribution over topics. That is, choose one of many multinomial distributions, each which mixes the topics in different proportions.



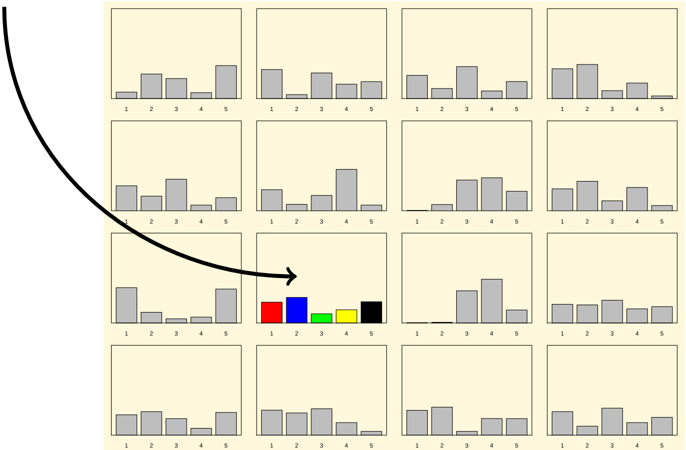
First Part

Randomly choose a distribution over topics. That is, choose one of many multinomial distributions, each which mixes the topics in different proportions.



First Part

Randomly choose a distribution over topics. That is, choose one of many multinomial distributions, each which mixes the topics in different proportions.



Second Part

Then, for every word in the document. . .

Second Part

Then, for every word in the document...

1. Randomly choose a topic from the distribution over topics from step 1.

Second Part

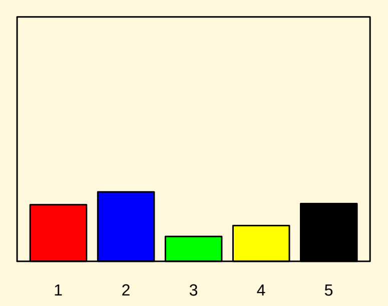
Then, for every word in the document...

1. Randomly choose a topic from the distribution over topics from step 1.
2. Randomly choose a word from the distribution over the vocabulary that the topic implies.

Second Part

Then, for every word in the document...

1. Randomly choose a topic from the distribution over topics from step 1.
2. Randomly choose a word from the distribution over the vocabulary that the topic implies.

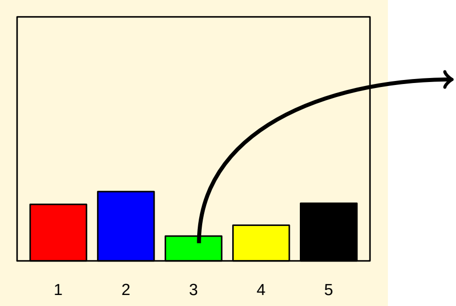


Second Part

Then, for every word in the document...

1. Randomly choose a topic from the distribution over topics from step 1.

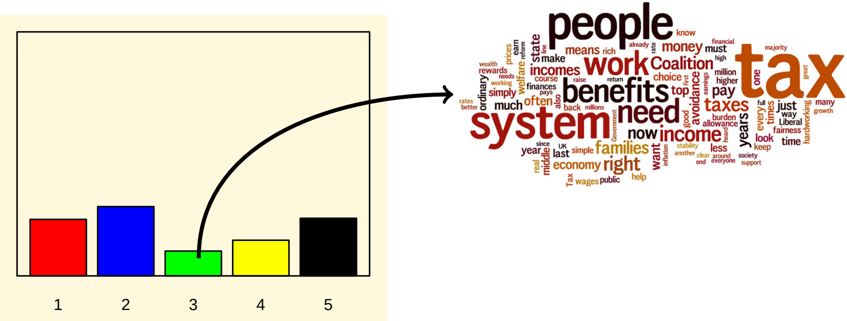
2. Randomly choose a word from the distribution over the vocabulary that the topic implies.



Second Part

Then, for every word in the document...

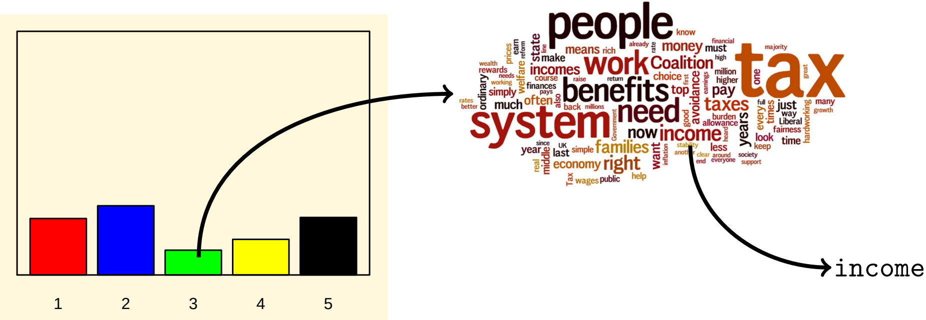
1. Randomly choose a topic from the distribution over topics from step 1.
2. Randomly choose a word from the distribution over the vocabulary that the topic implies.



Second Part

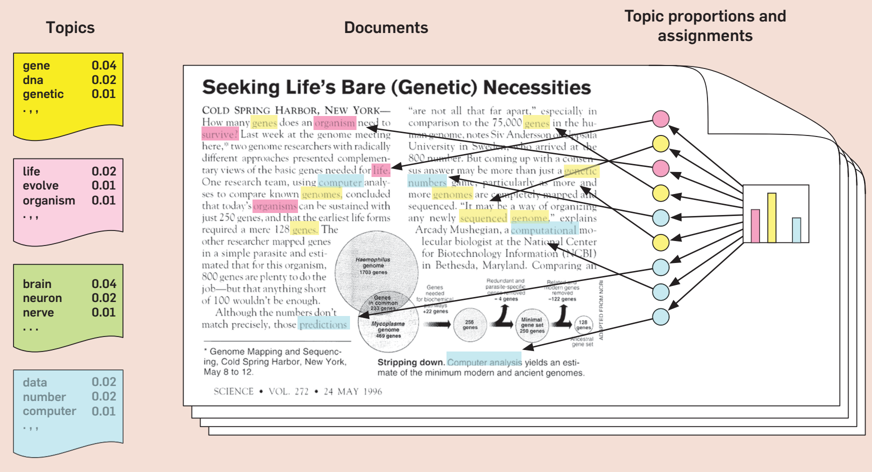
Then, for every word in the document...

1. Randomly choose a topic from the distribution over topics from step 1.
2. Randomly choose a word from the distribution over the vocabulary that the topic implies.



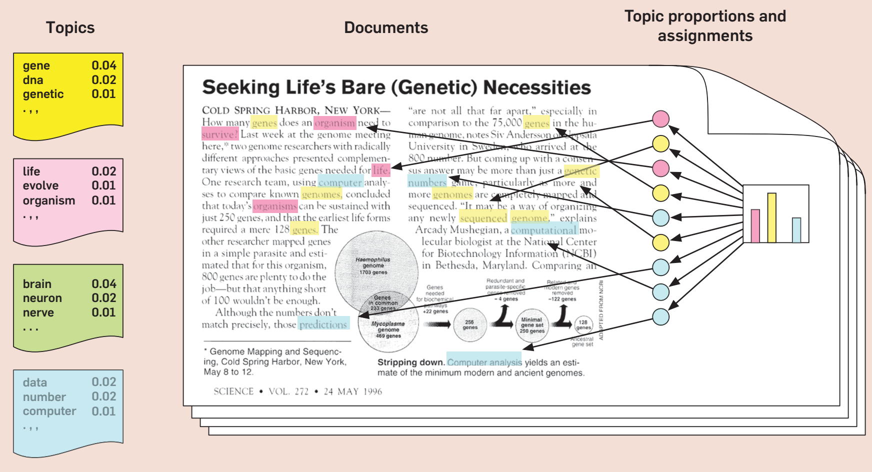
Topic Modeling a Document (Blei, 2012)

Topic Modeling a Document (Blei, 2012)



Note that all documents share same set of topics:

Topic Modeling a Document (Blei, 2012)



Note that all documents share same set of topics: but some (e.g. neuro) may be (basically) absent in a given document.

Notes

Notes

Some of our variables—the documents which contain the words—are observable.

Notes

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are $\boxed{\text{latent}}$.

Notes

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are $\boxed{\text{latent}}$.

We need a distribution from which to draw the per-document topic distribution. We use a $\boxed{\text{Dirichlet}}$ distribution as a prior for that.

Notes

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are latent.

We need a distribution from which to draw the per-document topic distribution. We use a Dirichlet distribution as a prior for that.

And Dirichlet is used for the allocation of the words in the documents to different topics:

Notes

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are $\boxed{\text{latent}}$.

We need a distribution from which to draw the per-document topic distribution. We use a $\boxed{\text{Dirichlet}}$ distribution as a prior for that.

And Dirichlet is used for the $\boxed{\text{allocation}}$ of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

Notes

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are $\boxed{\text{latent}}$.

We need a distribution from which to draw the per-document topic distribution. We use a $\boxed{\text{Dirichlet}}$ distribution as a prior for that.

And Dirichlet is used for the $\boxed{\text{allocation}}$ of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

→ Latent

Notes

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are $\boxed{\text{latent}}$.

We need a distribution from which to draw the per-document topic distribution. We use a $\boxed{\text{Dirichlet}}$ distribution as a prior for that.

And Dirichlet is used for the $\boxed{\text{allocation}}$ of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

→ Latent Dirichlet

Notes

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are $\boxed{\text{latent}}$.

We need a distribution from which to draw the per-document topic distribution. We use a $\boxed{\text{Dirichlet}}$ distribution as a prior for that.

And Dirichlet is used for the $\boxed{\text{allocation}}$ of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

$\rightarrow$ Latent Dirichlet Allocation.

Notes

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are $\boxed{\text{latent}}$.

We need a distribution from which to draw the per-document topic distribution. We use a $\boxed{\text{Dirichlet}}$ distribution as a prior for that.

And Dirichlet is used for the $\boxed{\text{allocation}}$ of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

$\rightarrow$ Latent Dirichlet Allocation. $\boxed{\textbf{LDA}}$.

A little more formally. . .

A little more formally. . .

LDA is a very popular topic model:

A little more formally...

LDA is a very popular topic model: a probabilistic procedure to generate topics. Thus, a 'generative' model.

A little more formally...

LDA is a very popular topic model: a probabilistic procedure to generate topics. Thus, a 'generative' model.

There are D documents in the corpus.

A little more formally...

LDA is a very popular topic model: a probabilistic procedure to generate topics. Thus, a 'generative' model.

There are D documents in the corpus. There are V terms in these D documents.

A little more formally...

LDA is a very popular topic model: a probabilistic procedure to generate topics. Thus, a 'generative' model.

There are D documents in the corpus. There are V terms in these D documents. For now suppose we know the K topic distributions: there are K multinomials containing V elements each.

A little more formally...

LDA is a very popular topic model: a probabilistic procedure to generate topics. Thus, a 'generative' model.

There are D documents in the corpus. There are V terms in these D documents. For now suppose we know the K topic distributions: there are K multinomials containing V elements each.

The multinomial distribution for the ith topic is denoted β_i, and $|\beta_i| = V$, meaning that the 'size' of this multinomial is equal to the number of different words in the corpus.

So, a little more formally. . .

So, a little more formally...

For each document...

So, a little more formally...

For each document...

1. Randomly choose a distribution over topics (multinomial of length K)

So, a little more formally...

For each document...

1. Randomly choose a distribution over topics (multinomial of length K)
2. Then, for every word in the document...

So, a little more formally...

For each document...

1. Randomly choose a distribution over topics (multinomial of length K)
2. Then, for every word in the document...
    1. Probabilistically draw one of the K topics from the distribution over topics from step 1. E.g. draw β_j

So, a little more formally...

For each document...

1. Randomly choose a distribution over topics (multinomial of length K)
2. Then, for every word in the document...
   1. Probabilistically draw one of the K topics from the distribution over topics from step 1. E.g. draw β_j
   2. Probabilistically draw one of the V words from β_j

Even more formally...

Even more formally. . .

For each document. . .

Even more formally...

For each document...

1. Draw a topic distribution $\theta_d \sim \mathrm{Dir}(\alpha)$, where $\mathrm{Dir}(\cdot)$ is a draw from a symmetric (uniform) Dirichlet distribution with scaling parameter, α

Even more formally...

For each document...

1. Draw a topic distribution $\theta_d \sim \text{Dir}(\alpha)$, where $\text{Dir}(\cdot)$ is a draw from a symmetric (uniform) Dirichlet distribution with scaling parameter, α

2. Then, for every word in the document...

Even more formally...

For each document...

1. Draw a topic distribution $\theta_d \sim \mathrm{Dir}(\alpha)$, where $\mathrm{Dir}(\cdot)$ is a draw from a symmetric (uniform) Dirichlet distribution with scaling parameter, α

2. Then, for every word in the document...

    1. Draw a specific topic $z_{d,n} \sim \mathrm{multi}(\theta_d)$ where $\mathrm{multi}(\cdot)$ is a multinomial. Here $z_{d,n}$ is the topic *assignment* for the word in the nth position of the dth document. E.g. word in position 2 in document 5 is from Topic 6.

Even more formally...

For each document...

1. Draw a topic distribution $\theta_d \sim \text{Dir}(\alpha)$, where $\text{Dir}(\cdot)$ is a draw from a symmetric (uniform) Dirichlet distribution with scaling parameter, α

2. Then, for every word in the document...

    1. Draw a specific topic $z_{d,n} \sim \text{multi}(\theta_d)$ where $\text{multi}(\cdot)$ is a multinomial. Here $z_{d,n}$ is the topic *assignment* for the word in the nth position of the dth document. E.g. word in position 2 in document 5 is from Topic 6. BTW, the multinomial has only one trial.

Even more formally...

For each document...

1. Draw a topic distribution $\theta_d \sim \text{Dir}(\alpha)$, where $\text{Dir}(\cdot)$ is a draw from a symmetric (uniform) Dirichlet distribution with scaling parameter, α

2. Then, for every word in the document...

    1. Draw a specific topic $z_{d,n} \sim \text{multi}(\theta_d)$ where $\text{multi}(\cdot)$ is a multinomial. Here $z_{d,n}$ is the topic *assignment* for the word in the nth position of the dth document. E.g. word in position 2 in document 5 is from Topic 6. BTW, the multinomial has only one trial.

    2. Draw a word $w_{d,n} \sim \beta_{z_{d,n}}$. Here, $w_{d,n}$ is the word in the nth position of the dth document and it is being drawn from topic $\beta_{z_{d,n}}$.

Even more formally...

For each document...

1. Draw a topic distribution $\theta_d \sim \text{Dir}(\alpha)$, where $\text{Dir}(\cdot)$ is a draw from a symmetric (uniform) Dirichlet distribution with scaling parameter, α
2. Then, for every word in the document...
   1. Draw a specific topic $z_{d,n} \sim \text{multi}(\theta_d)$ where $\text{multi}(\cdot)$ is a multinomial. Here $z_{d,n}$ is the topic *assignment* for the word in the nth position of the dth document. E.g. word in position 2 in document 5 is from Topic 6. BTW, the multinomial has only one trial.
   2. Draw a word $w_{d,n} \sim \beta_{z_{d,n}}$. Here, $w_{d,n}$ is the word in the nth position of the dth document and it is being drawn from topic $\beta_{z_{d,n}}$. E.g. word in position 2 in document 5 is from Topic 6 and turns out to be 'income' in this particular case.

Aside: Dirichlet distribution

Aside: Dirichlet distribution

The Dirichlet distribution is a conjugate prior for the multinomial ('categorical' if you only have one trial) distribution.

Aside: Dirichlet distribution

The Dirichlet distribution is a conjugate prior for the multinomial ('categorical' if you only have one trial) distribution. Makes certain calculations easier.

Aside: Dirichlet distribution

The Dirichlet distribution is a conjugate prior for the multinomial ('categorical' if you only have one trial) distribution. Makes certain calculations easier.

It is parameterized by a vector of positive real numbers $\boldsymbol{\alpha}$. In principle, one can have $\alpha_1, \ldots, \alpha_k$ be different concentration parameters, but LDA uses special symmetric Dirichlet where all the values of $\boldsymbol{\alpha}$ are the same.

Aside: Dirichlet distribution

The Dirichlet distribution is a conjugate prior for the multinomial ('categorical' if you only have one trial) distribution. Makes certain calculations easier.

It is parameterized by a vector of positive real numbers $\boldsymbol{\alpha}$. In principle, one can have $\alpha_1, \ldots, \alpha_k$ be different concentration parameters, but LDA uses special symmetric Dirichlet where all the values of $\boldsymbol{\alpha}$ are the same.

Larger values of α (assuming we are in symmetric case) mean we think (*a priori*) that documents are generally an even mix of the topics.

Aside: Dirichlet distribution

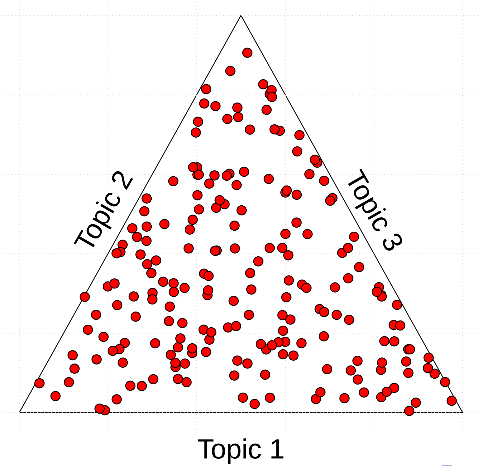
The Dirichlet distribution is a conjugate prior for the multinomial ('categorical' if you only have one trial) distribution. Makes certain calculations easier.

It is parameterized by a vector of positive real numbers $\boldsymbol{\alpha}$. In principle, one can have $\alpha_1, \ldots, \alpha_k$ be different concentration parameters, but LDA uses special symmetric Dirichlet where all the values of $\boldsymbol{\alpha}$ are the same.

Larger values of α (assuming we are in symmetric case) mean we think (*a priori*) that documents are generally an even mix of the topics. If α is small (less than 1) we think a given document is generally from one or a few topics.

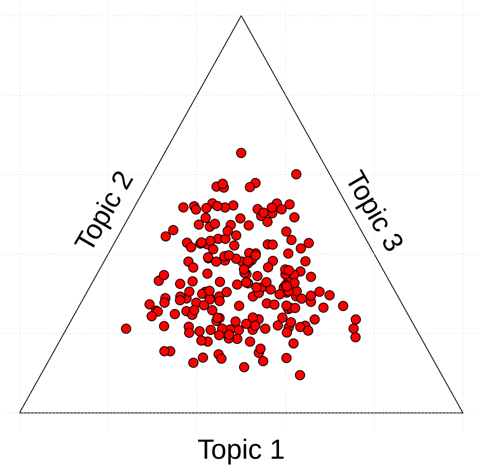
Example of Dirichlet

200 documents, 3 topics, $\alpha = 1$
(uniform)



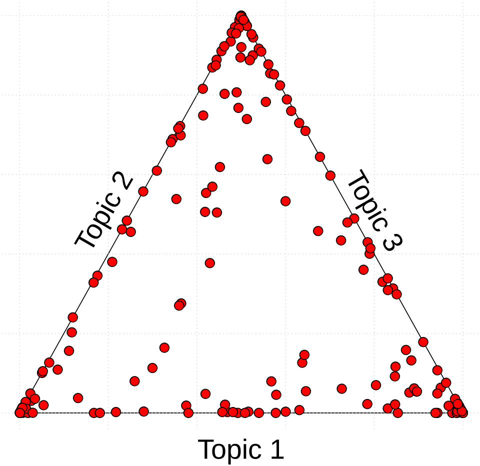
Example of Dirichlet

200 documents, 3 topics, $\alpha = 5$



Example of Dirichlet

200 documents, 3 topics, $\alpha = 0.2$



And actually...

And actually...

We also use a symmetric Dirichlet prior on the per topic word distributions.

And actually. . .

We also use a symmetric Dirichlet prior on the per topic word distributions. That is, the prior on the β_is.

And actually. . .

We also use a symmetric Dirichlet prior on the per topic word distributions. That is, the prior on the β_is.

$\rightarrow$ A high concentration parameter means each topic is a mixture of most of the words.

And actually...

We also use a symmetric Dirichlet prior on the per topic word distributions. That is, the prior on the β_is.

$\rightarrow$ A high concentration parameter means each topic is a mixture of most of the words. A low concentration parameter means each topic is a mixture of a few of the words.

And actually...

We also use a symmetric Dirichlet prior on the per topic word distributions. That is, the prior on the β_is.

$\rightarrow$ A high concentration parameter means each topic is a mixture of most of the words. A low concentration parameter means each topic is a mixture of a few of the words.

In practice, one can estimate the concentration parameters, or simply set them at suggested values.

And actually...

We also use a symmetric Dirichlet prior on the per topic word distributions. That is, the prior on the β_is.

$\rightarrow$ A high concentration parameter means each topic is a mixture of most of the words. A low concentration parameter means each topic is a mixture of a few of the words.

In practice, one can estimate the concentration parameters, or simply set them at suggested values.

We want topic models to be similar as we increase number of topics. Can use asymmetric priors for per-document topic distributions (the θs).

And actually...

We also use a symmetric Dirichlet prior on the per topic word distributions. That is, the prior on the β_is.

$\rightarrow$ A high concentration parameter means each topic is a mixture of most of the words. A low concentration parameter means each topic is a mixture of a few of the words.

In practice, one can estimate the concentration parameters, or simply set them at suggested values.

We want topic models to be similar as we increase number of topics. Can use asymmetric priors for per-document topic distributions (the θs). Asymmetric priors on per-topic word distributions don't do much.

And actually...

We also use a symmetric Dirichlet prior on the per topic word distributions. That is, the prior on the β_is.

$\rightarrow$ A high concentration parameter means each topic is a mixture of most of the words. A low concentration parameter means each topic is a mixture of a few of the words.

In practice, one can estimate the concentration parameters, or simply set them at suggested values.

We want topic models to be similar as we increase number of topics. Can use asymmetric priors for per-document topic distributions (the θs). Asymmetric priors on per-topic word distributions don't do much. Wallach et al "Rethinking LDA: Why Priors Matter"

We now know that...

We now know that…

We observe $w_{d,n}$.

We now know that...

We observe $w_{d,n}$. And there are N words in a given document.

We now know that…

We observe $w_{d,n}$. And there are N words in a given document. And D documents in a corpus.

We now know that...

We observe $w_{d,n}$. And there are N words in a given document. And D documents in a corpus.

The $w_{d,n}$ we see depends on $z_{d,n}$, the topic assignment for that word ("this word will be from topic 4").

We now know that...

We observe $w_{d,n}$. And there are N words in a given document. And D documents in a corpus.

The $w_{d,n}$ we see depends on $z_{d,n}$, the topic assignment for that word ("this word will be from topic 4").

The $z_{d,n}$ depends on θ_d, the topic mix for a given document d in which the words sit.

We now know that...

We observe $w_{d,n}$. And there are N words in a given document. And D documents in a corpus.

The $w_{d,n}$ we see depends on $z_{d,n}$, the topic assignment for that word ("this word will be from topic 4").

The $z_{d,n}$ depends on θ_d, the topic mix for a given document d in which the words sit.

The θ_d depends on our prior for the relevant Dirichlet,

We now know that...

We observe $w_{d,n}$. And there are N words in a given document. And D documents in a corpus.

The $w_{d,n}$ we see depends on $z_{d,n}$, the topic assignment for that word ("this word will be from topic 4").

The $z_{d,n}$ depends on θ_d, the topic mix for a given document d in which the words sit.

The θ_d depends on our prior for the relevant Dirichlet, α.

We now know that...

We observe $w_{d,n}$. And there are N words in a given document. And D documents in a corpus.

The $w_{d,n}$ we see depends on $z_{d,n}$, the topic assignment for that word ("this word will be from topic 4").

The $z_{d,n}$ depends on θ_d, the topic mix for a given document d in which the words sit.

The θ_d depends on our prior for the relevant Dirichlet, α.

And we know that the actual value that $w_{d,n}$ takes depends on the distribution over words that the relevant topic entails, the β ("the word from topic 4 is "`income`" in this case")

While the β depends on the prior for the relevant Dirichlet, η

Plate Diagram for LDA

Plate Diagram for LDA

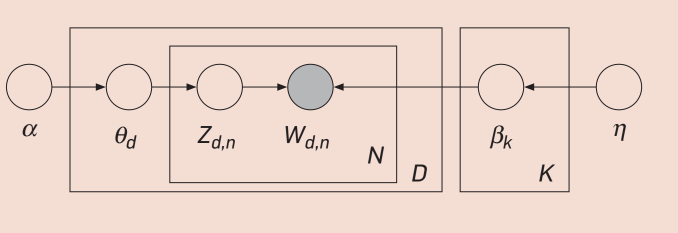
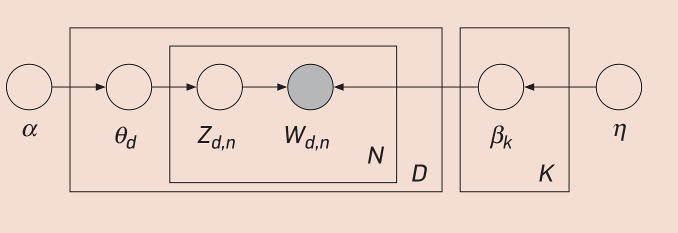
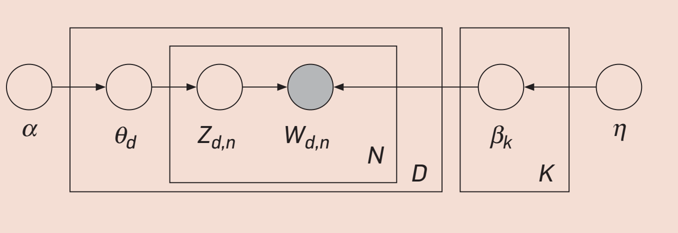


Plate Diagram for LDA



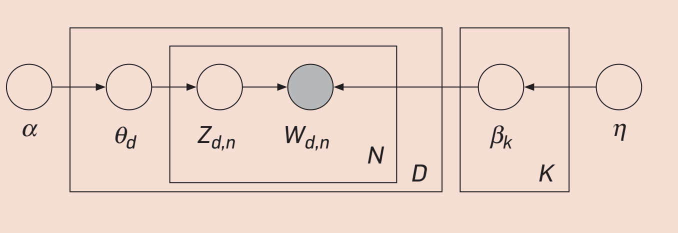
Solid nodes are observed;

Plate Diagram for LDA



Solid nodes are observed; empty nodes are latent.

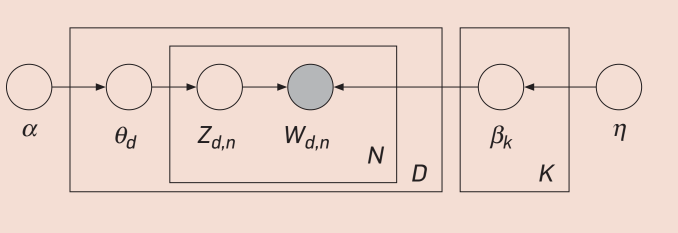
Plate Diagram for LDA



Solid nodes are observed; empty nodes are latent.

Plates imply replication.

Plate Diagram for LDA



Solid nodes are observed; empty nodes are latent.

Plates imply replication.

Note that $w_{d,n}$ depends on $z_{d,n}$ (the mix of topics for that document) and $\beta_{1:K}$ (all the topics in terms of their distributions over the words).

Bayesian Inference: Crash Course/Reminder

Recall that. . .

Recall that...

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

Recall that...

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

So,

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

And, Bayes Theorem tell us that

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$$

Conditional Inference Using Bayes' Theorem

Conditional Inference Using Bayes' Theorem

What is the probability that Republicans will win in 2024?

Conditional Inference Using Bayes' Theorem

What is the probability that Republicans will win in 2024?

We might have a guess at this.

Conditional Inference Using Bayes' Theorem

What is the probability that Republicans will win in 2024?

We might have a guess at this. Summarize as a $\boxed{\text{prior}}$, $\Pr(A)$.

Conditional Inference Using Bayes' Theorem

What is the probability that Republicans will win in 2024?

We might have a guess at this. Summarize as a $\boxed{\text{prior}}$, $\Pr(A)$.

e.g. I think it is Bernoulli with $p = 0.8$ (so, likely, but by no means certain). You could have $\mathcal{B}(\frac{2}{3})$ instead.

Conditional Inference Using Bayes' Theorem

What is the probability that Republicans will win in 2024?

We might have a guess at this. Summarize as a $\boxed{\text{prior}}$, $\Pr(A)$.

e.g. I think it is Bernoulli with $p = 0.8$ (so, likely, but by no means certain). You could have $\mathcal{B}(\frac{2}{3})$ instead.

Presumably, we think the *actual* probability GOP win depends on a whole set of things:

Conditional Inference Using Bayes' Theorem

What is the probability that Republicans will win in 2024?

We might have a guess at this. Summarize as a $\boxed{\text{prior}}$, $\Pr(A)$.

e.g. I think it is Bernoulli with $p = 0.8$ (so, likely, but by no means certain). You could have $\mathcal{B}(\frac{2}{3})$ instead.

Presumably, we think the *actual* probability GOP win depends on a whole set of things: state of economy, success in war, etc.

Conditional Inference Using Bayes' Theorem

What is the probability that Republicans will win in 2024?

We might have a guess at this. Summarize as a $\boxed{\text{prior}}$, $\Pr(A)$.

e.g. I think it is Bernoulli with $p = 0.8$ (so, likely, but by no means certain). You could have $\mathcal{B}(\frac{2}{3})$ instead.

Presumably, we think the *actual* probability GOP win depends on a whole set of things: state of economy, success in war, etc. Call these data B.

Conditional Inference Using Bayes' Theorem

What is the probability that Republicans will win in 2024?

We might have a guess at this. Summarize as a $\boxed{\text{prior}}$, $\Pr(A)$.

e.g. I think it is Bernoulli with $p = 0.8$ (so, likely, but by no means certain). You could have $\mathcal{B}(\frac{2}{3})$ instead.

Presumably, we think the *actual* probability GOP win depends on a whole set of things: state of economy, success in war, etc. Call these data B.

Notice

Notice

Interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)} = \Pr(\text{win}|\text{data})$.

Notice

Interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)} = \Pr(\text{win}|\text{data})$.

This is the estimated probability of winning,

Notice

Interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)} = \Pr(\text{win}|\text{data})$.

This is the estimated probability of winning, given observed data. It is equal to

Notice

Interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)} = \Pr(\text{win}|\text{data})$.

This is the estimated probability of winning, given observed data. It is equal to the product of the prior beliefs about how that probability might be distributed

Notice

Interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)} = \Pr(\text{win}|\text{data})$.

This is the estimated probability of winning, given observed data. It is equal to the product of the prior beliefs about how that probability might be distributed $\times$ $\Pr(B|A)$ ... divided by the (unconditional) probability of the data.

Notice

Interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)} = \Pr(\text{win}|\text{data})$.

This is the estimated probability of winning, given observed data. It is equal to the product of the prior beliefs about how that probability might be distributed $\times$ $\Pr(B|A)$ ...divided by the (unconditional) probability of the data.

We will get at $Pr(B|A)$ via a model for the DGP.

Notice

Interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)} = \Pr(\text{win}|\text{data})$.

This is the estimated probability of winning, given observed data. It is equal to the product of the prior beliefs about how that probability might be distributed $\times$ $\Pr(B|A)$ … divided by the (unconditional) probability of the data.

We will get at $Pr(B|A)$ via a model for the DGP.

We will call $\Pr(A|B)$ the posterior.

Notice

Interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)} = \Pr(\text{win}|\text{data})$.

This is the estimated probability of winning, given observed data. It is equal to the product of the prior beliefs about how that probability might be distributed $\times$ $\Pr(B|A)$ ... divided by the (unconditional) probability of the data.

We will get at $Pr(B|A)$ via a model for the DGP.

We will call $\Pr(A|B)$ the posterior. It will imply that some values are more plausible than others,

Notice

Interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)} = \Pr(\text{win}|\text{data})$.

This is the estimated probability of winning, given observed data. It is equal to the product of the prior beliefs about how that probability might be distributed $\times$ $\Pr(B|A)$ ... divided by the (unconditional) probability of the data.

We will get at $Pr(B|A)$ via a model for the DGP.

We will call $\Pr(A|B)$ the posterior. It will imply that some values are more plausible than others, and we might want various features of it, like the mean or median. NB: in MLE, we would report the mode of $\boldsymbol{\theta}$

More formally

More formally

We have data $\mathbf{X}$.

More formally

We have data $\mathbf{X}$. We want to make an inference from it,

More formally

We have data $\mathbf{X}$. We want to make an inference from it, so we assume it was produced by some (topic) model M, which has parameters $\boldsymbol{\theta}$.

More formally

We have data $\mathbf{X}$. We want to make an inference from it, so we assume it was produced by some (topic) model M, which has parameters $\boldsymbol{\theta}$.

Further, assume that $\mathbf{X}$ are iid, and generated by some likelihood $p(\mathbf{X}|\boldsymbol{\theta}, M)$.

More formally

We have data $\mathbf{X}$. We want to make an inference from it, so we assume it was produced by some (topic) model M, which has parameters $\boldsymbol{\theta}$.

Further, assume that $\mathbf{X}$ are iid, and generated by some likelihood $p(\mathbf{X}|\boldsymbol{\theta}, M)$.

The posterior over the parameters = likelihood of the data, conditioned on particular values for the parameters, multiplied by our prior.

More formally

We have data $\mathbf{X}$. We want to make an inference from it, so we assume it was produced by some (topic) model M, which has parameters $\boldsymbol{\theta}$.

Further, assume that $\mathbf{X}$ are iid, and generated by some likelihood $p(\mathbf{X}|\boldsymbol{\theta}, M)$.

The posterior over the parameters = likelihood of the data, conditioned on particular values for the parameters, multiplied by our prior.

Then: $\Pr(A|B) = \boxed{\dfrac{\Pr(A)\Pr(B|A)}{\Pr(B)}}$.

Then: $\Pr(A|B) = \boxed{\dfrac{\Pr(A)\Pr(B|A)}{\Pr(B)}}$.

Now:

Then: $\Pr(A|B) = \boxed{\dfrac{\Pr(A)\Pr(B|A)}{\Pr(B)}}$.

Now:

$$p(\boldsymbol{\theta}|\mathbf{X}, M) = \frac{p(\mathbf{X}, \boldsymbol{\theta}|M)}{\int p(\mathbf{X}, \boldsymbol{\theta}|M) d\boldsymbol{\theta}} = \boxed{\frac{p(\theta|M)p(\mathbf{X}|\boldsymbol{\theta}, M)}{p(\mathbf{X}|M)}}$$

We refer to denominator as the 'normalizing constant' or the 'marginal likelihood'—

Then: $\Pr(A|B) = \boxed{\dfrac{\Pr(A)\Pr(B|A)}{\Pr(B)}}.$

Now:

$$p(\boldsymbol{\theta}|\mathbf{X}, M) = \frac{p(\mathbf{X}, \boldsymbol{\theta}|M)}{\int p(\mathbf{X}, \boldsymbol{\theta}|M) d\boldsymbol{\theta}} = \boxed{\frac{p(\theta|M) p(\mathbf{X}|\boldsymbol{\theta}, M)}{p(\mathbf{X}|M)}}$$

We refer to denominator as the 'normalizing constant' or the 'marginal likelihood'—because it's the likelihood with the model parameters integrated out.

Then: $\Pr(A|B) = \boxed{\dfrac{\Pr(A)\Pr(B|A)}{\Pr(B)}}$.

Now:

$$p(\boldsymbol{\theta}|\mathbf{X}, M) = \frac{p(\mathbf{X}, \boldsymbol{\theta}|M)}{\int p(\mathbf{X}, \boldsymbol{\theta}|M)d\boldsymbol{\theta}} = \boxed{\frac{p(\theta|M)p(\mathbf{X}|\boldsymbol{\theta}, M)}{p(\mathbf{X}|M)}}$$

We refer to denominator as the 'normalizing constant' or the 'marginal likelihood'—because it's the likelihood with the model parameters integrated out.

NB: sometimes called the evidence in the Bayesian context, and is integral of numerator over support of θ (weighted by how plausible each value is)

Tasks

Tasks

1. <u>Predicting</u> values of new data points $\mathbf{X}_{\text{new}}$ given the observed data.

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
   $p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
   $p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$
   But, this integral is often intractable because posterior, $p(\boldsymbol{\theta}|\mathbf{X}, M)$ is very high dimensional or has awkward form.

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
   $p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$
   But, this integral is often intractable because posterior, $p(\boldsymbol{\theta}|\mathbf{X}, M)$ is very high dimensional or has awkward form.

2. Marginalization:

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
   $p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$
   But, this integral is often intractable because posterior, $p(\boldsymbol{\theta}|\mathbf{X}, M)$ is very high dimensional or has awkward form.

2. Marginalization: may have $\boldsymbol{\theta}$ and another 'nuisance' parameter $\mathbf{z}$,

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
   $p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$
   But, this integral is often intractable because posterior, $p(\boldsymbol{\theta}|\mathbf{X}, M)$ is very high dimensional or has awkward form.

2. Marginalization: may have $\boldsymbol{\theta}$ and another 'nuisance' parameter $\mathbf{z}$, and we want $p(\boldsymbol{\theta}|\mathbf{X}, M) = \int p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{X}, M)d\mathbf{z}$.

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
   $p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$
   But, this integral is often intractable because posterior, $p(\boldsymbol{\theta}|\mathbf{X}, M)$ is very high dimensional or has awkward form.

2. Marginalization: may have $\boldsymbol{\theta}$ and another 'nuisance' parameter $\mathbf{z}$, and we want $p(\boldsymbol{\theta}|\mathbf{X}, M) = \int p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{X}, M)d\mathbf{z}$. Which may require intractable integration.

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
   $p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$
   But, this integral is often intractable because posterior, $p(\boldsymbol{\theta}|\mathbf{X}, M)$ is very high dimensional or has awkward form.

2. Marginalization: may have $\boldsymbol{\theta}$ and another 'nuisance' parameter $\mathbf{z}$, and we want $p(\boldsymbol{\theta}|\mathbf{X}, M) = \int p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{X}, M)d\mathbf{z}$. Which may require intractable integration.

3. Model selection:

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
$p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$
But, this integral is often intractable because posterior, $p(\boldsymbol{\theta}|\mathbf{X}, M)$ is very high dimensional or has awkward form.

2. Marginalization: may have $\boldsymbol{\theta}$ and another 'nuisance' parameter $\mathbf{z}$, and we want $p(\boldsymbol{\theta}|\mathbf{X}, M) = \int p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{X}, M)d\mathbf{z}$. Which may require intractable integration.

3. Model selection: might want to compare models, and see which is most plausible.

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
   $p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$
   But, this integral is often intractable because posterior, $p(\boldsymbol{\theta}|\mathbf{X}, M)$ is very high dimensional or has awkward form.

2. Marginalization: may have $\boldsymbol{\theta}$ and another 'nuisance' parameter $\mathbf{z}$, and we want $p(\boldsymbol{\theta}|\mathbf{X}, M) = \int p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{X}, M)d\mathbf{z}$. Which may require intractable integration.

3. Model selection: might want to compare models, and see which is most plausible. But that requires calculating $\int p(\mathbf{X}, \boldsymbol{\theta}|M)d\boldsymbol{\theta}$

Tasks

1. Predicting values of new data points $\mathbf{X}_{\text{new}}$ given the observed data. We have to average over the posterior:
   $p(\mathbf{X}_{\text{new}}|\mathbf{X}, M) = \int p(\mathbf{X}_{\text{new}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{X}, M)d\boldsymbol{\theta}$
   But, this integral is often intractable because posterior, $p(\boldsymbol{\theta}|\mathbf{X}, M)$ is very high dimensional or has awkward form.

2. Marginalization: may have $\boldsymbol{\theta}$ and another 'nuisance' parameter $\mathbf{z}$, and we want $p(\boldsymbol{\theta}|\mathbf{X}, M) = \int p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{X}, M)d\mathbf{z}$. Which may require intractable integration.

3. Model selection: might want to compare models, and see which is most plausible. But that requires calculating $\int p(\mathbf{X}, \boldsymbol{\theta}|M)d\boldsymbol{\theta}$ ('evidence') which may be intractable.

So. . .

So. . .

Bayesian approaches require integration.

So. . .

Bayesian approaches require integration. But those integrals may be difficult.

So. . .

Bayesian approaches require integration. But those integrals may be difficult.

Instead, we can approximate. Two main ideas:

So. . .

Bayesian approaches require integration. But those integrals may be difficult.

Instead, we can approximate. Two main ideas:

Stochastic: we can use Monte Carlo simulation methods to sample from the posterior distribution, e.g. $p(\boldsymbol{\theta}|\mathbf{X}, M)$.

So. . .

Bayesian approaches require integration. But those integrals may be difficult.

Instead, we can approximate. Two main ideas:

Stochastic: we can use Monte Carlo simulation methods to sample from the posterior distribution, e.g. $p(\boldsymbol{\theta}|\mathbf{X}, M)$. Markov chain Monte Carlo is a way to do this,

So. . .

Bayesian approaches require integration. But those integrals may be difficult.

Instead, we can approximate. Two main ideas:

Stochastic: we can use Monte Carlo simulation methods to sample from the posterior distribution, e.g. $p(\boldsymbol{\theta}|\mathbf{X}, M)$. Markov chain Monte Carlo is a way to do this, with the Gibbs Sampler a specific example.

So. . .

Bayesian approaches require integration. But those integrals may be difficult.

Instead, we can approximate. Two main ideas:

Stochastic: we can use Monte Carlo simulation methods to sample from the posterior distribution, e.g. $p(\boldsymbol{\theta}|\mathbf{X}, M)$. Markov chain Monte Carlo is a way to do this, with the Gibbs Sampler a specific example.

Deterministic: most notably for topic models, variational inference.

So. . .

Bayesian approaches require integration. But those integrals may be difficult.

Instead, we can approximate. Two main ideas:

Stochastic: we can use Monte Carlo simulation methods to sample from the posterior distribution, e.g. $p(\boldsymbol{\theta}|\mathbf{X}, M)$. Markov chain Monte Carlo is a way to do this, with the Gibbs Sampler a specific example.

Deterministic: most notably for topic models, variational inference. Idea is to write down posterior as product of well-known distributions.

So...

Bayesian approaches require integration. But those integrals may be difficult.

Instead, we can approximate. Two main ideas:

Stochastic: we can use Monte Carlo simulation methods to sample from the posterior distribution, e.g. $p(\boldsymbol{\theta}|\mathbf{X}, M)$. Markov chain Monte Carlo is a way to do this, with the Gibbs Sampler a specific example.

Deterministic: most notably for topic models, variational inference. Idea is to write down posterior as product of well-known distributions. This will approximate the true posterior (use KL divergence to get as close as possible to it).

Hmm. . . What's so great about Bayesian Methods?

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE.

Hmm…What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

Hmm. . . What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy:

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1 Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense.

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1 Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with p-values and notions of repeating (finite) phenomena.

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with *p*-values and notions of repeating (finite) phenomena. We have priors from previous work

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1 Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with p-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process.

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with p-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process. Automatic handling of missing data.

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with p-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process. Automatic handling of missing data. More sensible handling of uncertainty.

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with p-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process. Automatic handling of missing data. More sensible handling of uncertainty. Data overwhelms prior.

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with p-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process. Automatic handling of missing data. More sensible handling of uncertainty. Data overwhelms prior.

2. Practicality:

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with p-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process. Automatic handling of missing data. More sensible handling of uncertainty. Data overwhelms prior.
2. Practicality: for many problems, MLE doesn't work well.

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with *p*-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process. Automatic handling of missing data. More sensible handling of uncertainty. Data overwhelms prior.

2. Practicality: for many problems, MLE doesn't work well. Perhaps because there are lots of parameters, but not much data:

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with p-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process. Automatic handling of missing data. More sensible handling of uncertainty. Data overwhelms prior.
2. Practicality: for many problems, MLE doesn't work well. Perhaps because there are lots of parameters, but not much data: priors can help here.

Hmm. . . What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with *p*-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process. Automatic handling of missing data. More sensible handling of uncertainty. Data overwhelms prior.

2. Practicality: for many problems, MLE doesn't work well. Perhaps because there are lots of parameters, but not much data: priors can help here. Or perhaps because the model, like multinomial probit, involves evaluating something complicated, analytically:

Hmm...What's so great about Bayesian Methods?

Bayesian computation is difficult and intensive compared to, say, MLE. So why do it?

1. Philosophy: Bayesian (v frequentist) approach to probability and inference simply makes more sense. We do away with p-values and notions of repeating (finite) phenomena. We have priors from previous work and can include them explicitly as part of scientific process. Automatic handling of missing data. More sensible handling of uncertainty. Data overwhelms prior.

2. Practicality: for many problems, MLE doesn't work well. Perhaps because there are lots of parameters, but not much data: priors can help here. Or perhaps because the model, like multinomial probit, involves evaluating something complicated, analytically: Bayesian methods can arbitrarily approximate the integral.

Crash course complete: back to LDA

Estimation

Estimation

Ultimately,

Estimation

Ultimately, we will use the observed data, the words,

Estimation

Ultimately, we will use the observed data, the words, to make an inference about the latent parameters:

Estimation

Ultimately, we will use the observed data, the words, to make an inference about the latent parameters: the βs, the zs, the θs.

Estimation

Ultimately, we will use the observed data, the words, to make an inference about the latent parameters: the βs, the zs, the θs. That will be a conditional probability.

Estimation

Ultimately, we will use the observed data, the words, to make an inference about the latent parameters: the βs, the zs, the θs. That will be a conditional probability.

We start with the joint distribution implied by the problem:

Estimation

Ultimately, we will use the observed data, the words, to make an inference about the latent parameters: the βs, the zs, the θs. That will be a conditional probability.

We start with the joint distribution implied by the problem:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$$

Estimation

Ultimately, we will use the observed data, the words, to make an inference about the latent parameters: the βs, the zs, the θs. That will be a conditional probability.

We start with the joint distribution implied by the problem:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$$

$$\prod_{K}^{i=1} p(\beta_i) \prod_{D}^{d=1} p(\theta_d) \left(\prod_{N}^{n=1} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right)$$

Posterior

Posterior

Generally we want

$$p(\boldsymbol{\theta}|\mathbf{X}, M) = \frac{p(\mathbf{X}, \boldsymbol{\theta}|M)}{\int p(\mathbf{X}, \boldsymbol{\theta}|M) d\boldsymbol{\theta}} = \boxed{\frac{p(\theta|M) p(\mathbf{X}|\boldsymbol{\theta}, M)}{p(\mathbf{X}|M)}}$$

Posterior

Generally we want

$$p(\boldsymbol{\theta}|\mathbf{X}, M) = \frac{p(\mathbf{X}, \boldsymbol{\theta}|M)}{\int p(\mathbf{X}, \boldsymbol{\theta}|M)d\boldsymbol{\theta}} = \boxed{\frac{p(\theta|M)p(\mathbf{X}|\boldsymbol{\theta}, M)}{p(\mathbf{X}|M)}}$$

Here (Blei, 2012), that is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

Posterior

Generally we want

$$p(\boldsymbol{\theta}|\mathbf{X}, M) = \frac{p(\mathbf{X}, \boldsymbol{\theta}|M)}{\int p(\mathbf{X}, \boldsymbol{\theta}|M)d\boldsymbol{\theta}} = \boxed{\frac{p(\theta|M)p(\mathbf{X}|\boldsymbol{\theta}, M)}{p(\mathbf{X}|M)}}$$

Here (Blei, 2012), that is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

Can get ‘evidence’ (denominator) by summing joint distribution over every possible topic structure:

Posterior

Generally we want

$$p(\boldsymbol{\theta}|\mathbf{X}, M) = \frac{p(\mathbf{X}, \boldsymbol{\theta}|M)}{\int p(\mathbf{X}, \boldsymbol{\theta}|M)d\boldsymbol{\theta}} = \boxed{\frac{p(\theta|M)p(\mathbf{X}|\boldsymbol{\theta}, M)}{p(\mathbf{X}|M)}}$$

Here (Blei, 2012), that is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

Can get 'evidence' (denominator) by summing joint distribution over every possible topic structure: every possible way of assigning each word to a topic.

Posterior

Generally we want

$$p(\boldsymbol{\theta}|\mathbf{X}, M) = \frac{p(\mathbf{X}, \boldsymbol{\theta}|M)}{\int p(\mathbf{X}, \boldsymbol{\theta}|M)d\boldsymbol{\theta}} = \boxed{\frac{p(\theta|M)p(\mathbf{X}|\boldsymbol{\theta}, M)}{p(\mathbf{X}|M)}}$$

Here (Blei, 2012), that is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

Can get 'evidence' (denominator) by summing joint distribution over every possible topic structure: every possible way of assigning each word to a topic. But this is impossible, so simulate/approximate.

Results

Results

For a user-selected k, a typical implementation of LDA will return. . .

Results

For a user-selected k, a typical implementation of LDA will return...

The word distribution for each topic.

Results

For a user-selected k, a typical implementation of LDA will return...

The word distribution for each topic.

The topic distribution for each document.

Results

For a user-selected k, a typical implementation of LDA will return...

The word distribution for each topic.

The topic distribution for each document.

Some implementations allow you to estimate e.g. α, in which case this is also returned.

Results

For a user-selected k, a typical implementation of LDA will return...

The word distribution for each topic.

The topic distribution for each document.

Some implementations allow you to estimate e.g. α, in which case this is also returned. And perhaps some kind of fit statistic(s).

A Manifesto Example

A Manifesto Example

69 UK manifestos.

A Manifesto Example

69 UK manifestos. Some preprocessing.

A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics.

A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics. Has Gibbs sampling and variational options.

A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics. Has Gibbs sampling and variational options.

The (some selected) word distributions for each topic.

A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics. Has Gibbs sampling and variational options.

The (some selected) word distributions for each topic. Sum down the columns is one.

A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics. Has Gibbs sampling and variational options.

The (some selected) word distributions for each topic. Sum down the columns is one.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
conservative	0.00188	0.00088	0.00185	0.00221	0.00168
party	0.00145	0.00067	0.00066	0.00577	0.00093
general	0.00073	0.00033	0.00018	0.00192	0.00040
election	0.00079	0.00053	0.00022	0.00235	0.00076
manifesto	0.00059	0.00078	0.00032	0.00099	0.00048
⋮	⋮	⋮	⋮	⋮	⋮

Continued. . .

Continued. . .

'Top' 6 most frequent words in each topic:

Continued...

'Top' 6 most frequent words in each topic: might help interpretation (!)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	new	[markup]	new	must
2	local	government	people	labour	government
3	government	people	new	government	labour
4	new	continue	work	people	shall
5	tax	can	[markup]	shall	can
6	liberal	conservative	support	britain	policy

Continued...

'Top' 6 most frequent words in each topic: might help interpretation (!)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	new	[markup]	new	must
2	local	government	people	labour	government
3	government	people	new	government	labour
4	new	continue	work	people	shall
5	tax	can	[markup]	shall	can
6	liberal	conservative	support	britain	policy

Up to analyst to label the topics!

Continued...

'Top' 6 most frequent words in each topic: might help interpretation (!)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	new	[markup]	new	must
2	local	government	people	labour	government
3	government	people	new	government	labour
4	new	continue	work	people	shall
5	tax	can	[markup]	shall	can
6	liberal	conservative	support	britain	policy

Up to analyst to label the topics!

Meaningless 'junk' topics not unusual:

Continued...

'Top' 6 most frequent words in each topic: might help interpretation (!)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	new	[markup]	new	must
2	local	government	people	labour	government
3	government	people	new	government	labour
4	new	continue	work	people	shall
5	tax	can	[markup]	shall	can
6	liberal	conservative	support	britain	policy

Up to analyst to label the topics!

Meaningless 'junk' topics not unusual: debate as to whether one has to interpret every topic.

Continued

Continued

The topic distribution for each document. . .

Continued

The topic distribution for each document...

Continued

The topic distribution for each document. . .

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
doc 1	0.00009	0.00009	0.00009	0.00009	0.99965
doc 2	0.00011	0.00011	0.00011	0.00011	0.99954
doc 3	0.00010	0.00010	0.00010	0.00010	0.99959
doc 4	0.00006	0.00006	0.00006	0.00006	0.99978
doc 5	0.00002	0.00002	0.00002	0.00002	0.99991
doc 6	0.00019	0.00019	0.00019	0.00019	0.99924
⋮	⋮	⋮	⋮	⋮	⋮

Continued

The topic distribution for each document...

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
doc 1	0.00009	0.00009	0.00009	0.00009	0.99965
doc 2	0.00011	0.00011	0.00011	0.00011	0.99954
doc 3	0.00010	0.00010	0.00010	0.00010	0.99959
doc 4	0.00006	0.00006	0.00006	0.00006	0.99978
doc 5	0.00002	0.00002	0.00002	0.00002	0.99991
doc 6	0.00019	0.00019	0.00019	0.00019	0.99924
⋮	⋮	⋮	⋮	⋮	⋮

Practical Notes I

Practical Notes I

Texts are usually preprocessed:

Practical Notes I

Texts are usually preprocessed: stop words removed,

Practical Notes I

Texts are usually preprocessed: stop words removed, (very) rare tokens removed.

Practical Notes I

Texts are usually preprocessed: stop words removed, (very) rare tokens removed. Punctuation often removed.

Practical Notes I

Texts are usually preprocessed: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

Practical Notes I

Texts are usually preprocessed: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the number of topics, K, is not picked automatically.

Practical Notes I

Texts are usually preprocessed: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the number of topics, K, is not picked automatically. Analysts select various Ks and check that their results are ‘robust’. But see over.

Practical Notes I

Texts are usually preprocessed: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the number of topics, K, is not picked automatically. Analysts select various Ks and check that their results are 'robust'. But see over.

As with all unsupervised learning,

Practical Notes I

Texts are usually preprocessed: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the number of topics, K, is not picked automatically. Analysts select various Ks and check that their results are 'robust'. But see over.

As with all unsupervised learning, interpretation is non-trivial, and requires a lot of validation. Rant: 'just-so' stories abound. Lazy analysts conclude whatever they want.

Practical Notes I

Texts are usually preprocessed: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the number of topics, K, is not picked automatically. Analysts select various Ks and check that their results are 'robust'. But see over.

As with all unsupervised learning, interpretation is non-trivial, and requires a lot of validation. Rant: 'just-so' stories abound. Lazy analysts conclude whatever they want.

Practical Notes II: Picking k

Practical Notes II: Picking k

Crudely: in social science,

Practical Notes II: Picking k

Crudely: in social science, researchers fit ‘enough’ topics until they see what they think they should.

Practical Notes II: Picking k

Crudely: in social science, researchers fit 'enough' topics until they see what they think they should. E.g. a certain topic—like `finance` suddenly peels off—so stop there.

Practical Notes II: Picking k

Crudely: in social science, researchers fit ‘enough’ topics until they see what they think they should. E.g. a certain topic—like `finance` suddenly peels off—so stop there.

$\rightarrow$ Check findings are robust in the neighborhood: if best model has $k = 35$, check $k = 30 - 40$ yields similar inferences.

Practical Notes II: Picking k

Crudely: in social science, researchers fit 'enough' topics until they see what they think they should. E.g. a certain topic—like `finance` suddenly peels off—so stop there.

$\rightarrow$ Check findings are robust in the neighborhood: if best model has $k = 35$, check $k = 30 - 40$ yields similar inferences.

NB: social scientists typically fit far fewer topics than CS, even to same data.

Practical Notes II: Picking k

Crudely: in social science, researchers fit ‘enough’ topics until they see what they think they should. E.g. a certain topic—like `finance` suddenly peels off—so stop there.

$\rightarrow$ Check findings are robust in the neighborhood: if best model has $k = 35$, check $k = 30 - 40$ yields similar inferences.

NB: social scientists typically fit far fewer topics than CS, even to same data.

Picking k, continued...

Picking k, continued...

CS: split into training and test sets.

Picking k, continued...

CS: split into training and test sets. In the training set,

Picking k, continued...

CS: split into training and test sets. In the training set,

1. pick some value of k and fit a topic model.

Picking k, continued...

CS: split into training and test sets. In the training set,

1. pick some value of k and fit a topic model.
2. record value of α (hyperparameter on document specific topic distributions) and word distributions for the topics (the βs)

Picking k, continued...

CS: split into training and test sets. In the training set,

1. pick some value of k and fit a topic model.
2. record value of α (hyperparameter on document specific topic distributions) and word distributions for the topics (the βs)

We'll write the βs as $\boldsymbol{\beta}$, then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\boldsymbol{\beta}, \alpha) = \sum_d \log p(w_d|\boldsymbol{\beta}, \alpha)$$

Picking k, continued. . .

CS: split into training and test sets. In the training set,

1. pick some value of k and fit a topic model.
2. record value of α (hyperparameter on document specific topic distributions) and word distributions for the topics (the βs)

We'll write the βs as $\boldsymbol{\beta}$, then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\boldsymbol{\beta}, \alpha) = \sum_d \log p(w_d|\boldsymbol{\beta}, \alpha)$$

where **w** are the words in the test set.

Picking k, continued...

CS: split into training and test sets. In the training set,

1. pick some value of k and fit a topic model.
2. record value of α (hyperparameter on document specific topic distributions) and word distributions for the topics (the βs)

We'll write the βs as $\boldsymbol{\beta}$, then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\boldsymbol{\beta}, \alpha) = \sum_d \log p(w_d|\boldsymbol{\beta}, \alpha)$$

where $\mathbf{w}$ are the words in the test set. Higher $\mathcal{L}$ implies better model.

Picking k, continued...

CS: split into training and test sets. In the training set,

1. pick some value of k and fit a topic model.
2. record value of α (hyperparameter on document specific topic distributions) and word distributions for the topics (the βs)

We'll write the βs as $\boldsymbol{\beta}$, then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\boldsymbol{\beta}, \alpha) = \sum_d \log p(w_d|\boldsymbol{\beta}, \alpha)$$

where $\mathbf{w}$ are the words in the test set. Higher $\mathcal{L}$ implies better model. Intuition is to calculate likelihood of seeing the test words, given what we know produced the training set.

Picking k, continued...

CS: split into training and test sets. In the training set,

1. pick some value of k and fit a topic model.
2. record value of α (hyperparameter on document specific topic distributions) and word distributions for the topics (the βs)

We'll write the βs as $\boldsymbol{\beta}$, then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\boldsymbol{\beta}, \alpha) = \sum_d \log p(w_d|\boldsymbol{\beta}, \alpha)$$

where $\mathbf{w}$ are the words in the test set. Higher $\mathcal{L}$ implies better model. Intuition is to calculate likelihood of seeing the test words, given what we know produced the training set.

Do this for all k.

Picking k, continued...

CS: split into training and test sets. In the training set,

1. pick some value of k and fit a topic model.
2. record value of α (hyperparameter on document specific topic distributions) and word distributions for the topics (the βs)

We'll write the βs as $\boldsymbol{\beta}$, then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\boldsymbol{\beta}, \alpha) = \sum_d \log p(w_d|\boldsymbol{\beta}, \alpha)$$

where $\mathbf{w}$ are the words in the test set. Higher $\mathcal{L}$ implies better model. Intuition is to calculate likelihood of seeing the test words, given what we know produced the training set.

Do this for all k.

In practice...

In practice. . .

Perplexity is popular option

In practice...

Perplexity is popular option

$$\text{perplexity} = \exp\left(-\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}}\right),$$

In practice...

Perplexity is popular option

$$\text{perplexity} = \exp\left(-\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}}\right),$$

where lower is better.

In practice...

Perplexity is popular option

$$\text{perplexity} = \exp\left(-\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}}\right),$$

where lower is better.

In general, $\mathcal{L}(\mathbf{w})$ is intractable,

In practice. . .

Perplexity is popular option

$$\text{perplexity} = \exp\left(-\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}}\right),$$

where lower is better.

In general, $\mathcal{L}(\mathbf{w})$ is intractable, but there are ways to approximate it.

In practice...

Perplexity is popular option

$$\text{perplexity} = \exp\left(-\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}}\right),$$

where lower is better.

In general, $\mathcal{L}(\mathbf{w})$ is intractable, but there are ways to approximate it.

But:

In practice...

Perplexity is popular option

$$\text{perplexity} = \exp\left(-\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}}\right),$$

where lower is better.

In general, $\mathcal{L}(\mathbf{w})$ is intractable, but there are ways to approximate it.

But: the topic models that hold-out calculations suggest are optimal and not much liked by humans!

In practice. . .

Perplexity is popular option

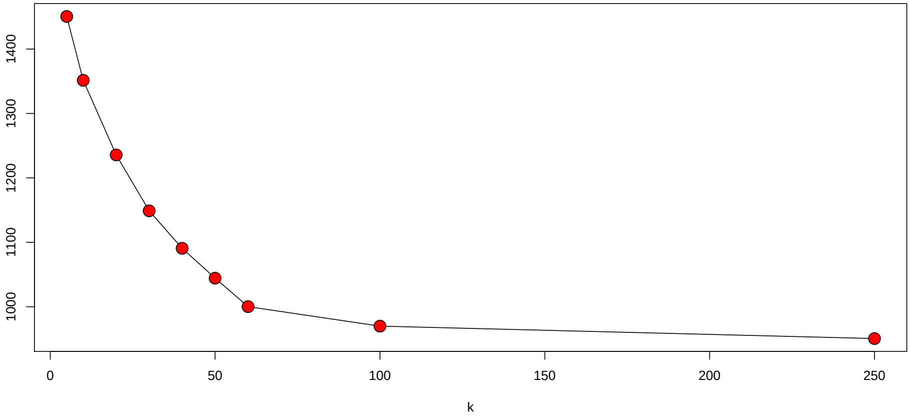
$$\text{perplexity} = \exp\left(-\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}}\right),$$

where lower is better.

In general, $\mathcal{L}(\mathbf{w})$ is intractable, but there are ways to approximate it.

But: the topic models that hold-out calculations suggest are optimal and not much liked by humans! "Reading Tea Leaves: How Humans Interpret Topic Models" by Chang et al.

Perplexity Likes a Lot of Topics (manifestos)



Pork to Policy (Catalinac, 2016)

Pork to Policy (Catalinac, 2016)

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case:

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy.

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area.

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

1. Rise of China?

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

1. Rise of China? Need to focus on security.

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

1. Rise of China? Need to focus on security.

vs.

2. Change in Electoral System?

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

1. Rise of China? Need to focus on security.

vs.

2. Change in Electoral System? Moved from promising pork to having to deliver policy as part of Westminster-style polity.

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

1. Rise of China? Need to focus on security.

vs.

2. Change in Electoral System? Moved from promising pork to having to deliver policy as part of Westminster-style polity.

To decide, we need data source that covers all lower house legislators

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

1. Rise of China? Need to focus on security.

vs.

2. Change in Electoral System? Moved from promising pork to having to deliver policy as part of Westminster-style polity.

To decide, we need data source that covers all lower house legislators where they set out their policy priorities over time.

Pork to Policy (Catalinac, 2016)

Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

1. Rise of China? Need to focus on security.

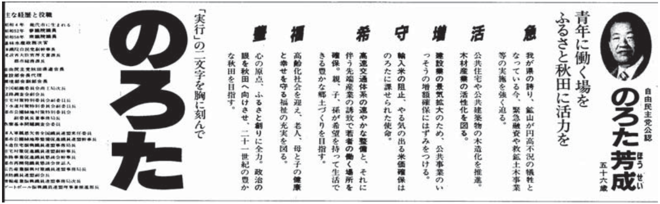
vs.

2. Change in Electoral System? Moved from promising pork to having to deliver policy as part of Westminster-style polity.

To decide, we need data source that covers all lower house legislators where they set out their policy priorities over time. See if/when they shift priorities.

Manifestos

Manifestos



自由民主党公認
のろた芳成（ほうせい）
五十六歳

青年に働く場を
ふるさと秋田に活力を

急
我が県の誇り、鉱山が円高不況の犠牲となっている今、緊急融資や救鉱土木事業等の実施を強く迫る。

活
公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。

増
建設業の景気拡大のため、公共事業のいっそうの増額確保にはずみをつける。

守
輸入米の阻止、やる気の出る米価確保はのろたに課せられた使命。

希
高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親、子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

福
高齢化社会を迎え、老人、母と子の健康と幸せを守る福祉の充実を図る。

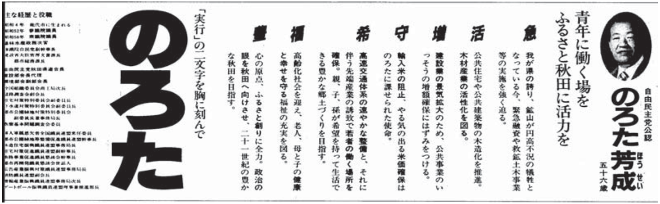
豊
心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

「実行」の二文字を胸に刻んで

のろた

主な経歴と役職

Manifestos



自由民主党公認
のろた芳成（ほうせい）
五十六歳

青年に働く場を
ふるさと秋田に活力を

急
我が県の誇り、鉱山が円高不況の犠牲となっている今、緊急融資や救鉱土木事業等の実施を強く迫る。

活
公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。

増
建設業の景気拡大のため、公共事業のいっそうの増額確保にはずみをつける。

守
輸入米の阻止、やる気の出る米価確保はのろたに課せられた使命。

希
高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親、子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

福
高齢化社会を迎え、老人、母と子の健康と幸せを守る福祉の充実を図る。

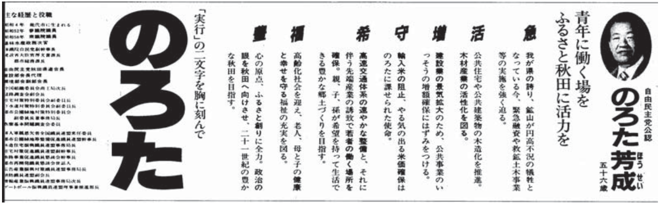
豊
心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

「実行」の二文字を胸に刻んで

のろた

主な経歴と役職

7,497.



自由民主党公認
のろた芳成（ほうせい）
五十六歳

青年に働く場を
ふるさと秋田に活力を

急
我が県の誇り、鉱山が円高不況の犠牲となっている今、緊急融資や救鉱土木事業等の実施を強く迫る。

活
公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。

増
建設業の景気拡大のため、公共事業のいっそうの増額確保にはずみをつける。

守
輸入米の阻止、やる気の出る米価確保はのろたに課せられた使命。

希
高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親、子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

福
高齢化社会を迎え、老人、母と子の健康と幸せを守る福祉の充実を図る。

豊
心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

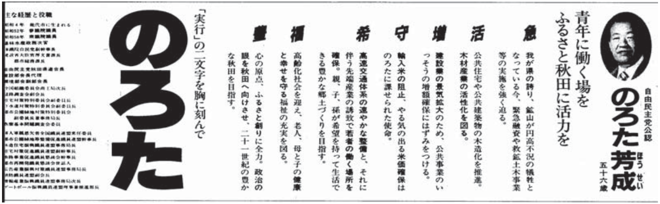
「実行」の二文字を胸に刻んで

のろた

主な経歴と役職

7,497. 1986–2009.

Manifestos



自由民主党公認

のろた芳成（ほうせい）

五十六歳

青年に働く場を

ふるさと秋田に活力を

急　我が県の誇り、鉱山が円高不況の犠牲となっている今、緊急融資や救鉱土木事業等の実施を強く迫る。

活　公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。

増　建設業の景気拡大のため、公共事業のいっそうの増額確保にはずみをつける。

守　輸入米の阻止、やる気の出る米価確保はのろたに課せられた使命。

希　高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親、子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

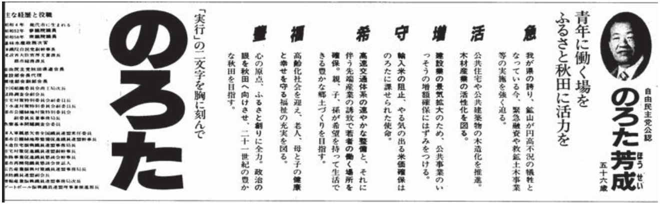
福　高齢化社会を迎え、老人、母と子の健康と幸せを守る福祉の充実を図る。

豊　心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

のろた

「実行」の二文字を胸に刻んで

7,497. 1986–2009. Standardized form.



自由民主党公認

のろた芳成(ほうせい)

五十六歳

青年に働く場を

ふるさと秋田に活力を

急

我が県の誇り、鉱山が円高不況の犠牲となっている今、緊急融資や救鉱土木事業等の実施を強く迫る。

活

公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。

増

建設業の景気拡大のため、公共事業のいっそうの増額確保にはずみをつける。

守

輸入米の阻止、やる気の出る米価確保はのろたに課せられた使命。

希

高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親、子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

福

高齢化社会を迎え、老人、母と子の健康と幸せを守る福祉の充実を図る。

心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

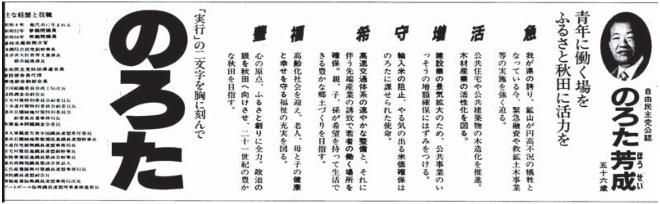
「実行」の二文字を胸に刻んで

のろた

7,497. 1986–2009. Standardized form.

"...instructed to write whatever they want in the form and return it before 5 PM of the first day of the campaign. At least two days before the election, local electoral commissions are required to distribute the forms of all candidates running in the district to all registered voters"

Manifestos



自由民主党公認
のろた芳成（ほうせい）
五十六歳

青年に働く場を
ふるさと秋田に活力を

急　我が県の誇り、鉱山が円高不況の犠牲となっている今、緊急融資や救鉱土木事業等の実施を強く迫る。

活　公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。

増　建設業の景気拡大のため、公共事業のいっそうの増額確保にはずみをつける。

守　輸入米の阻止、やる気の出る米価確保はのろたに課せられた使命。

希　高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親、子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

福　高齢化社会を迎え、老人、母と子の健康と幸せを守る福祉の充実を図る。

豊　心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

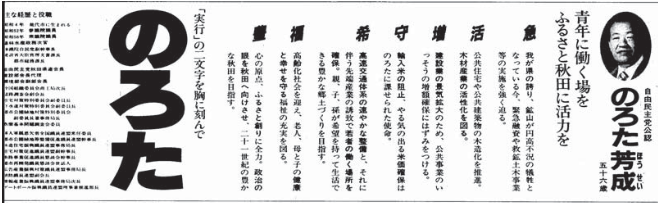
「実行」の二文字を胸に刻んで

のろた

7,497. 1986–2009. Standardized form.

“. . . instructed to write whatever they want in the form and return it before 5 PM of the first day of the campaign. At least two days before the election, local electoral commissions are required to distribute the forms of all candidates running in the district to all registered voters”

Manifestos were hand transcribed from microfilm.



自由民主党公認
のろた芳成（ほうせい）
五十六歳

青年に働く場を
ふるさと秋田に活力を

急
我が県の誇り、鉱山が円高不況の犠牲となっている今、緊急融資や救鉱土木事業等の実施を強く迫る。

活
公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。

増
建設業の景気拡大のため、公共事業のいっそうの増額確保にはずみをつける。

守
輸入米の阻止、やる気の出る米価確保はのろたに課せられた使命。

希
高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親、子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

福
高齢化社会を迎え、老人、母と子の健康と幸せを守る福祉の充実を図る。

豊
心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

「実行」の二文字を胸に刻んで

のろた

主な経歴と役職
昭和4年 能代市に生まれる
昭和52年 参議院議員
昭和58年 衆議院議員

7,497. 1986–2009. Standardized form.

“. . . instructed to write whatever they want in the form and return it before 5 PM of the first day of the campaign. At least two days before the election, local electoral commissions are required to distribute the forms of all candidates running in the district to all registered voters”

Manifestos were hand transcribed from microfilm. Japanese install of Windows/R used to fit LDA.

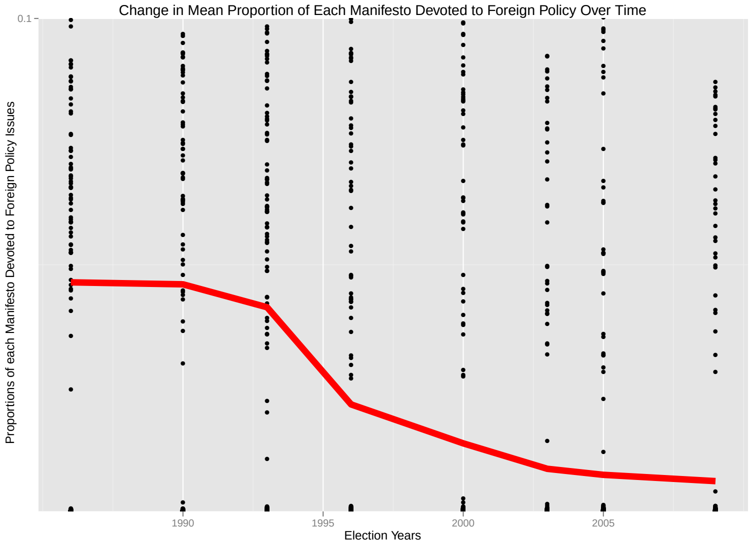
Topic Distribution over Words

Topic Distribution over Words

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	改革	年金	推進	区	政治	日本
2	郵政	円	整備	政策	改革	国
3	民営	廃止	図る	地域	国民	外交
4	小泉	改革	つとめる	まち	企業	国家
5	構造	兆	社会	鹿児島	自民党	社会
6	政府	実現	対策	全力	日本	国民
7	官	無駄	振興	選挙	共産党	保障
8	推進	日本	充実	国政	献金	安全
9	民	増税	促進	作り	金権	地域
10	自民党	削減	安定	横浜	党	拉致
11	日本	一元化	確立	対策	選挙	経済
12	制度	政権	企業	中小	禁止	守る
13	民間	子供	実現	発電	憲法	問題
14	年金	地域	中小	推進	腐敗	北朝鮮
15	実現	ひと	育成	エネルギー	団体	教育
16	進める	サラリーマン	制度	企業	区	責任
17	断行	制度	政治	声	ソ連	力
18	地方	議員	地域	実現	守る	創る
19	止める	金	福祉	活性	平和	安心
20	保障	民主党	事業	自民党	円	目指す
21	財政	年間	改革	地方	反対	誇り
22	作る	一掃	確保	尽くす	真	憲法
23	賛成	郵政	強化	商店	是正	可能
24	社会	道路	教育	いかす	一掃	道
25	国民	交代	施設	全国	悪政	未来
26	公務員	社会保険庁	生活	政党	抜本	ひと
27	力	月額	支援	ひと	定数	再生
28	経済	手当	環境	支援	政党	将来
29	国	談合	発展	経済	金丸	解決
30	安心	支援	施策	福祉	改憲	基本

Change in proportion of ‘Pork’ Topic

Change in proportion of ‘Pork’ Topic



Change in proportion of 'Foreign Policy' Topic

Change in proportion of ‘Foreign Policy’ Topic