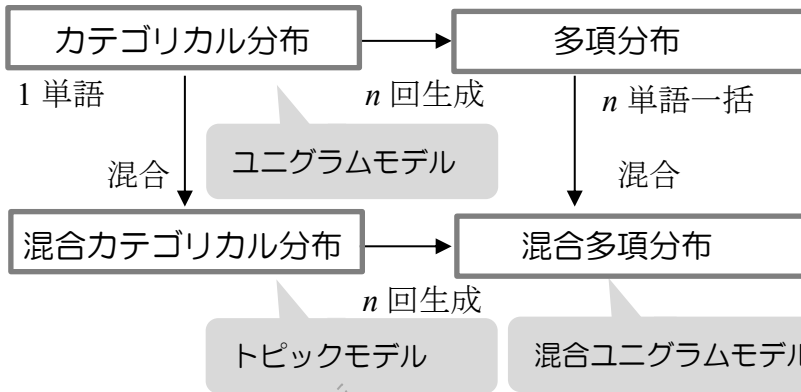


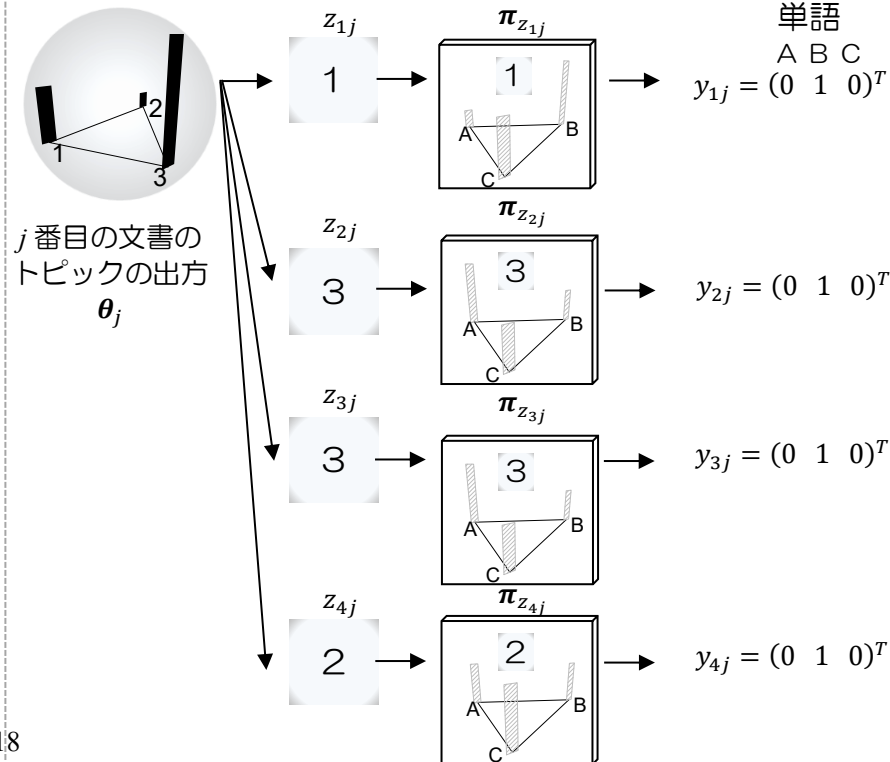
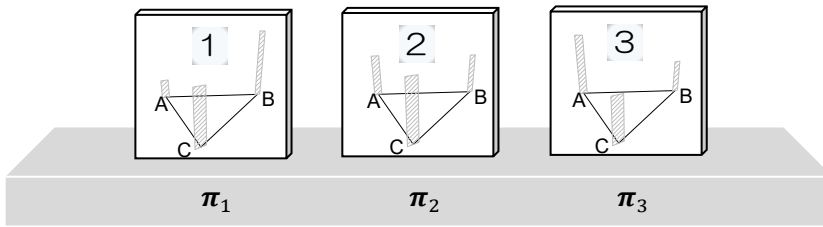
学びのポイント

- 二つ以上の分布がある混合比に基づいて混ぜられた分布のことを混合分布と呼ぶことが分かり、基本的な分布について視覚的に理解をすることができる。
- 混合分布の密度式は、「混ぜられる前の分布」×「その分布の混合比」を各分布に対して計算し、それらを足し合わせたものである、という点が理解できる。
- 混合正規分布、混合カテゴリカル分布、混合多項分布において、その密度式を理解でき、混合比を変化させると分布の形状がどのように変化するのか、説明することができる。
- 自然言語処理をはじめとする、文書への統計的アプローチでは、文の中の階層構造（統語構造）を捨象し、Bag-of-Words と呼ばれる表現で議論が行われていることが分かる。
- 各単語を、ベクトルとして表記するやり方が分かる。
- 各文書を、単語の集合としてベクトル表記するやり方が分かる。
- 文書の集合を、ベクトルとして表記するやり方が分かる。
- ユニグラムモデルが母集団に置いている仮定が理解でき、多項ロジスティック回帰において切片のみのモデルを仮定していることと同じであることを説明できる。
- 混合ユニグラムモデルが母集団に置いている仮定が理解でき、各単語が従う単語分布が決定される際に、潜在的なトピックという概念を用いて分布を混ぜていることを説明できる。
- トピックモデルが母集団に置いている過程が理解でき、混合ユニグラムモデルとは異なり、各文書にトピックが複数混ざって存在できる仕組みになっていることを説明できる。
- トピックモデルの推定結果について、具体的な事例に即して推論を行うことができ、潜在的なトピック（意味）の分析の仕方を説明できる。

見取り図



トピック1の単語分布 トピック2の単語分布 トピック3の単語分布



■ 目標

ここでは、「言語統計学 B」で登場したカテゴリカル分布／多項分布を発展させたユニグラムモデル、混合ユニグラムモデル、トピックモデルという統計モデルを扱います。

これらのモデルは、文書（あるいは文書における単語）の生成過程に関するモデルです。「生成過程」というのは、単語がどのような確率で選ばれて登場するのかというプロセスのことです。単純に考えれば、単語というものは、「犬」「猫」「科学」「恋人」…と、数こそ多いものの離散的なカテゴリーを形成していますから単語の出やすさを表すカテゴリカル分布を想定し、そこからランダムに選ばれているという想定を置きたくなります。

このシンプルな仮定のもと単語の生成過程をモデル化したものがユニグラムモデルです。これは、多項ロジスティック回帰の切片モデルであり、その構造の単純さからこの講義でもまずこのモデルに注目をしながら話を展開していきたいと考えています。

しかし、実際の研究の場面では、ユニグラムモデルよりも少し複雑な仮定を考えた方が切れの良い分析ができる場合もあります。例えば、単語の出やすさは、その文章のトピックに応じて変わるという仮定を置くこともできます。つまり、文学系の文書では「愛」とか「夫婦」とか「友情」といった単語が生成されやすいことでしょう。しかし、一方で生物の教科書では、「求愛」とか「つがい」とか「群れ」といった言葉が出る確率が高まっていると考えられます。このようにその文書がどのようなトピックに属するかで、カテゴリカル分布が変わるという想定を置いたものを混合ユニグラムモデルと呼びます。

しかし、この混合ユニグラムモデルよりも近年盛んに用いられるモデルとしてトピックモデルと呼ばれるモデルがあります。これは、各文書にトピックが一つと考える混合ユニグラムモデルとは対照的に、一つの文書の中に複数のトピックが混在していると考えられるモデルです。この講義の最終目的地は、みなさんにこのトピックモデルの考え方の基礎を、その応用事例とともに学んでもらうことにあります。

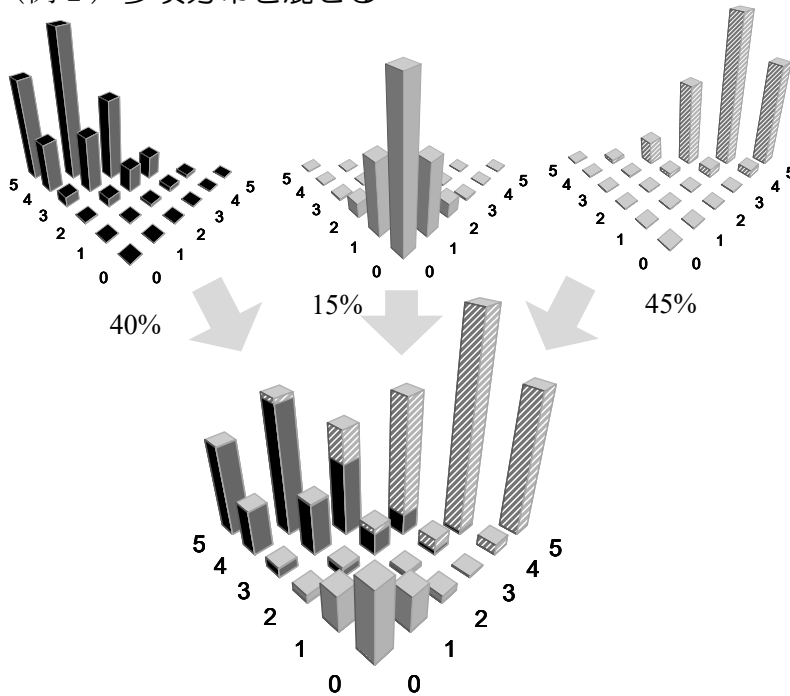
また、この混合ユニグラムモデルとトピックモデルのベースには二つ以上の分布を混ぜるという考え方が存在します。このような二つ以上の分布が混ざったものを混合分布と言い、話の前座として、この講義の前半では、混合分布と呼ばれるものがどのような性質を持つのか、視覚的な説明も交えながら、ゆっくりと解説をしていきます。

📖 ノート0 あらすじ：トピックモデルのねらい

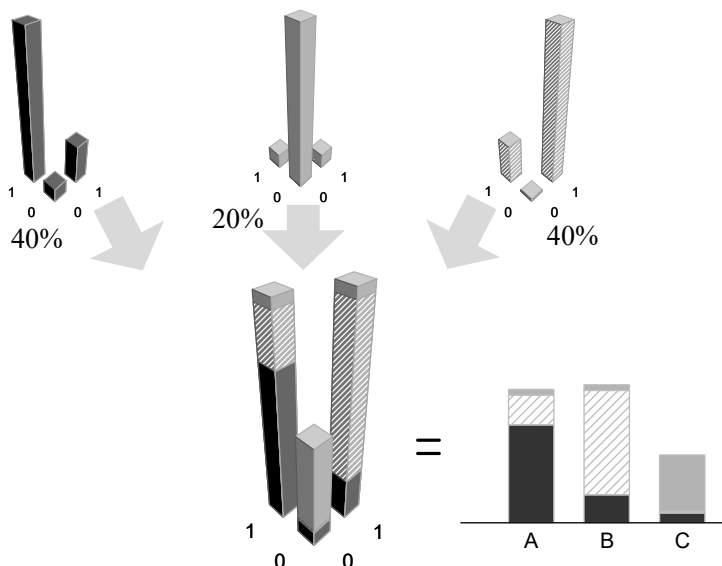
(1) モデル拡張の動機：分布の混合

複数の単純な事象が混ざったため複雑に見える事象がある。データから混ざり方と混ざる前の構造を推測したい。

(例1) 多項分布を混ぜる



(例2) カテゴリカル分布を混ぜる

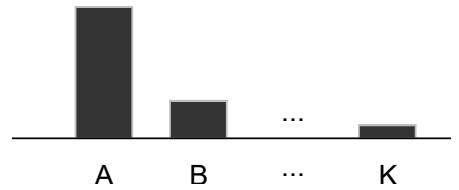


(2) トピックモデル

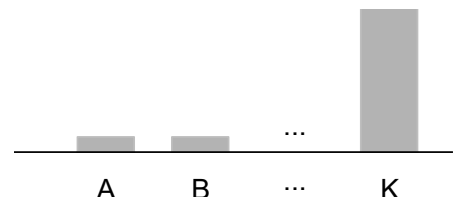
各「文書 j 」ごとに、混ざり方が異なるという仮定のもと、混ざり方（トピック分布）、混ざる前の構造（単語分布）を想定し、推定する統計モデル。

① 混ざる前の構造の分析：単語分布

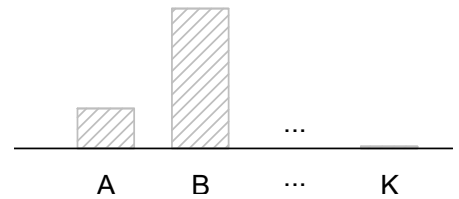
トピック1：



トピック2：

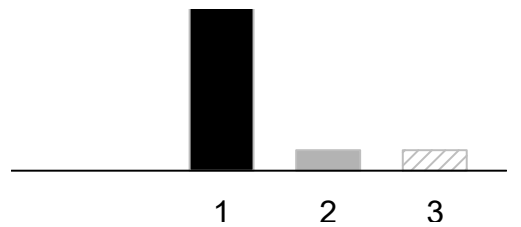


トピック3：

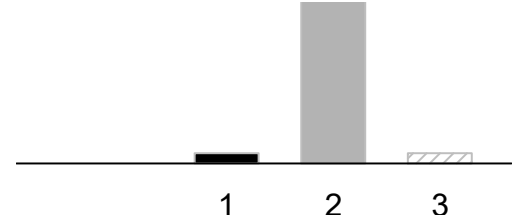


② 混ざり方の分析：トピック分布

文書1：

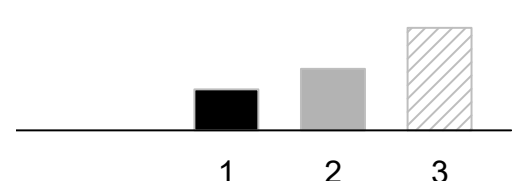


文書2：



⋮

文書J：



📖 ノート1 混合分布



正規分布に従うと仮定してきたけど、観測できない x ごとに正規分布になってるんじゃない？

(1) (有限) 混合モデル (Finite) Mixture Model

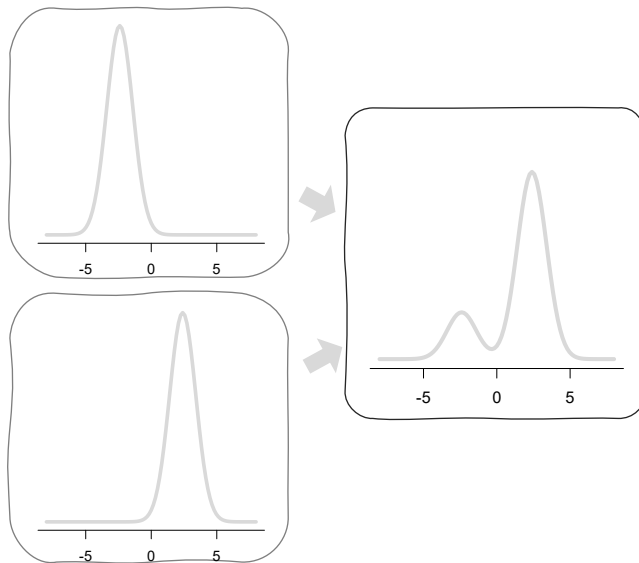
これは、(有限の) 複数のモデルをある割合 (混合比) に基づいて混ぜた確率モデルのこと。

(2) 連続型：混合正規分布

これは、(有限の) 複数の正規分布をある割合 (混合比) に基づいて混ぜた確率分布のこと。

(ケース1) 二個混ぜた場合

$$p(y|\theta, \mu, \sigma^2) = \theta_1 \times p(y|\mu_1, \sigma_1^2) + \theta_2 \times p(y|\mu_2, \sigma_2^2)$$



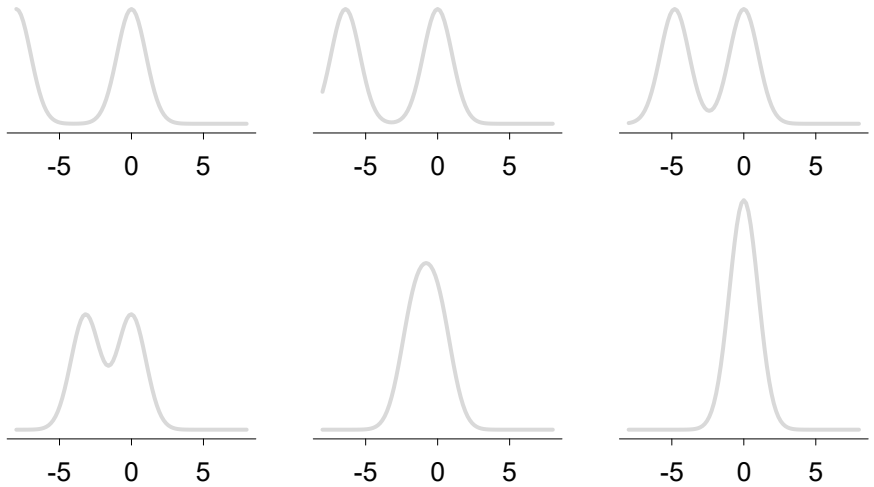
(ケース2) K 個混ぜた場合

$$p(y|\theta, \mu, \sigma^2) = \sum_{k=1}^K \theta_k \times p(y|\mu_k, \sigma_k^2)$$



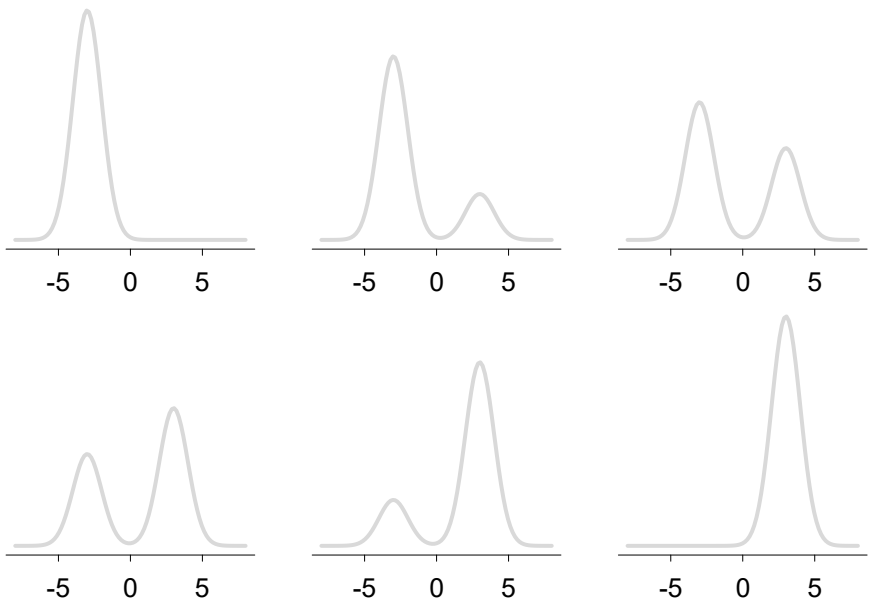
混合正規分布の例1：中心の位置をずらした例

$$0.5 \times p(y|\mu = k, \sigma^2 = 1) + 0.5 \times p(y|\mu = 0, \sigma^2 = 1)$$



混合正規分布の例2：混合比を変えた例

$$\theta_1 \times p(y|\mu = -3, \sigma^2 = 1) + \theta_2 \times p(y|\mu = 0, \sigma^2 = 1)$$

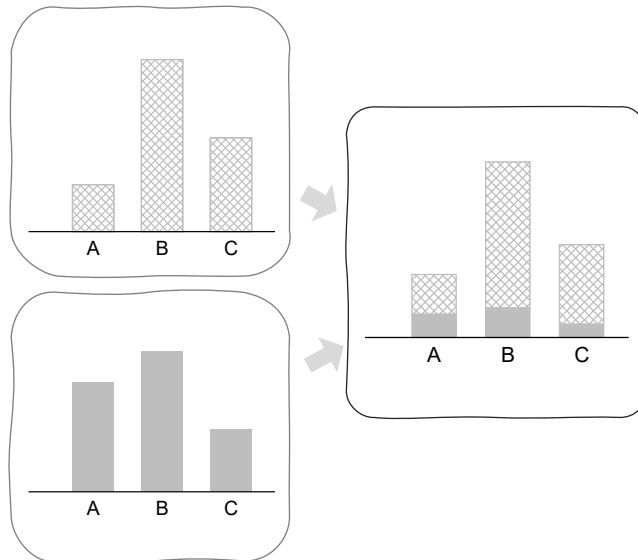


(3) 離散型 1：混合カテゴリカル分布

これは、(有限の) 複数のカテゴリカル分布をある割合 (混合比) に基づいて混ぜた確率分布のこと。

(ケース 1) 二個混ぜた場合

$$p(\mathbf{y}|\theta_1, \theta_2, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \theta_1 \times p(\mathbf{y}|\boldsymbol{\pi}_1) + \theta_2 \times p(\mathbf{y}|\boldsymbol{\pi}_2)$$



(ケース 2) K 個混ぜた場合

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^K \theta_k \times p(\mathbf{y}|\boldsymbol{\pi}_k)$$



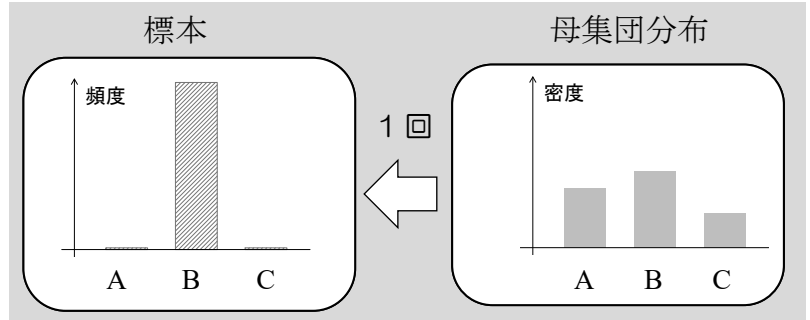
「単語データ」を従属変数にする

i 番目の単語を次のようにベクトルで表記する。

$$\mathbf{y}_i = \begin{matrix} & \text{A} & \text{B} & \text{C} \\ \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T & = & \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} & = & \begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \end{pmatrix} & \begin{matrix} \text{A} \\ \text{B} \\ \text{C} \end{matrix} \end{matrix}$$



密度関数 1 : カテゴリカル分布



① データ

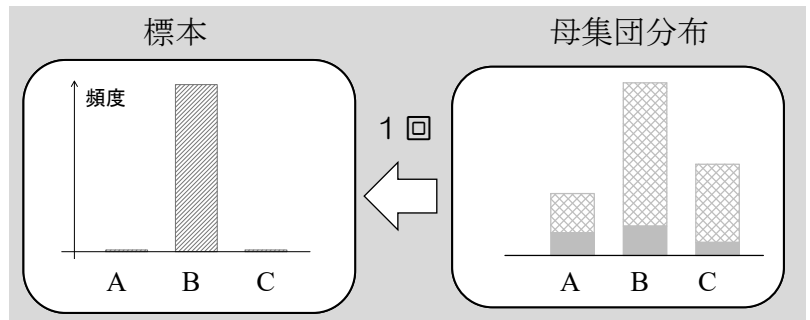
$$y_i = \begin{pmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{vi} \end{pmatrix}$$

② 確率密度関数

$$\begin{aligned} p(y_i | \boldsymbol{\pi}) &= \pi_1^{y_{1i}} \pi_2^{y_{2i}} \cdots \pi_v^{y_{vi}} \\ &= \prod_{v=1}^v \pi_v^{y_{vi}} \quad \cdots (\text{式 1}) \end{aligned}$$



密度関数 2 : 混合カテゴリカル分布



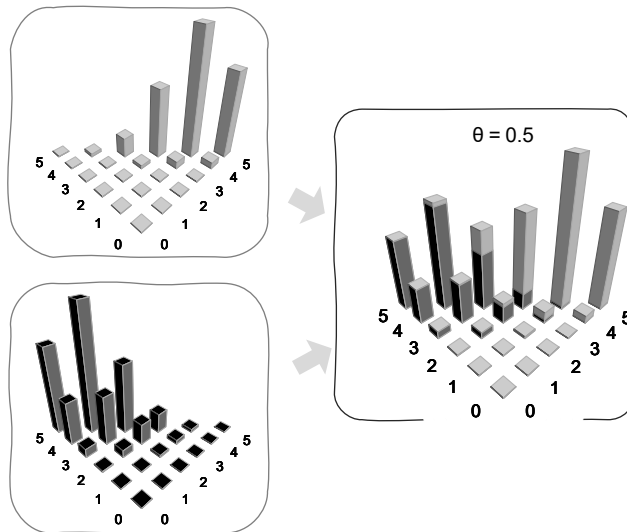
$$\begin{aligned} p(y_i | \boldsymbol{\theta}, \boldsymbol{\pi}) &= \theta_1 \times (\pi_{11}^{y_{1i}} \pi_{21}^{y_{2i}} \cdots \pi_{v1}^{y_{vi}}) \\ &\quad + \theta_2 \times (\pi_{12}^{y_{1i}} \pi_{22}^{y_{2i}} \cdots \pi_{v2}^{y_{vi}}) \\ &\quad + \cdots \\ &\quad + \theta_K \times (\pi_{1K}^{y_{1i}} \pi_{2K}^{y_{2i}} \cdots \pi_{vK}^{y_{vi}}) \\ &= \sum_{k=1}^K \theta_k \left(\prod_{v=1}^v \pi_{vk}^{y_{vi}} \right) \quad \cdots (\text{式 2}) \end{aligned}$$

(4) 離散型2：混合多項分布

これは、(有限の) 複数の多項分布をある割合 (混合比) に基づいて混ぜた確率分布のこと。

(ケース1) 二個混ぜた場合

$$p(\mathbf{y}|\theta_1, \theta_2, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, n) = \theta_1 \times p(\mathbf{y}|\boldsymbol{\pi}_1, n) + \theta_2 \times p(\mathbf{y}|\boldsymbol{\pi}_2, n)$$



(ケース2) K 個混ぜた場合

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\pi}, n) = \sum_{k=1}^K \theta_k \times p(\mathbf{y}|\boldsymbol{\pi}_k, n)$$



「文書データ」を従属変数にする

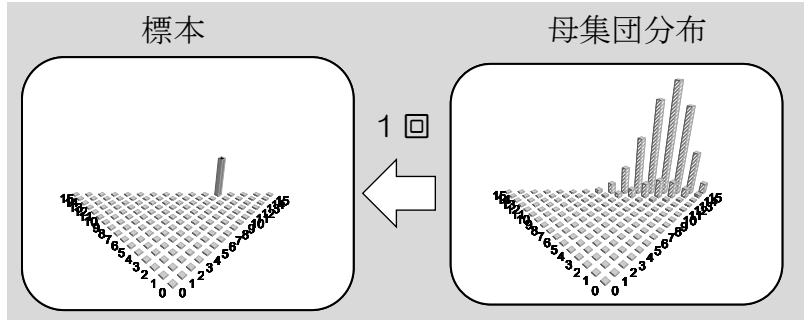
j 番目の文書 (N_j 個の単語の集合) をベクトルで表記する。

$$\mathbf{y}_j = (\mathbf{y}_{1j} \quad \mathbf{y}_{2j} \quad \cdots \quad \mathbf{y}_{N_j})^T = \begin{pmatrix} \mathbf{y}_{1j} \\ \mathbf{y}_{2j} \\ \vdots \\ \mathbf{y}_{ij} \\ \vdots \\ \mathbf{y}_{N_j} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{y}_{1j} \\ \mathbf{y}_{2j} \\ \vdots \\ (0 \quad 1 \quad \cdots \quad 0)^T \\ \vdots \\ \mathbf{y}_{N_j} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_{1j} \\ \mathbf{y}_{2j} \\ \vdots \\ (\mathbf{y}_{1ij} \quad \mathbf{y}_{2ij} \quad \cdots \quad \mathbf{y}_{vij} \quad \cdots \quad \mathbf{y}_{v_{ij}})^T \\ \vdots \\ \mathbf{y}_{N_j} \end{pmatrix}$$



密度関数 1 : 多項分布



① データ

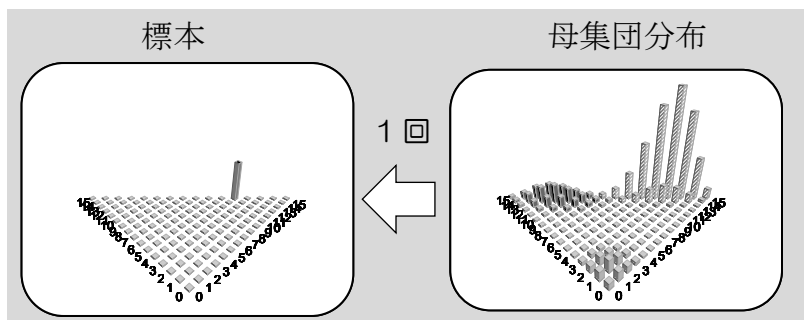
$$y_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{Vj} \end{pmatrix}$$

② 確率密度関数

$$\begin{aligned} p(\mathbf{y}_j | \boldsymbol{\pi}) &= \frac{N_j!}{N_{1j}! N_{2j}! \cdots N_{Vj}!} \times \prod_{i=1}^{N_j} \pi_1^{y_{1ij}} \pi_2^{y_{2ij}} \cdots \pi_V^{y_{Vij}} \\ &= (\text{定数}) \times \prod_{i=1}^{N_j} \prod_{v=1}^V \pi_v^{y_{vij}} \\ &= (\text{定数}) \times \prod_{v=1}^V \pi_v^{N_{vj}} \quad \dots (\text{式3}) \end{aligned}$$



密度関数 2 : 混合多項分布



$$p(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\pi}, n) = \sum_{k=1}^K \theta_k \times (\text{定数}) \times \prod_{v=1}^V \pi_v^{N_{vj}} \quad \dots (\text{式4})$$



(式3) の計算の詳細

$$\begin{aligned}
& \prod_{i=1}^{N_j} \pi_1^{y_{1ij}} \pi_2^{y_{2ij}} \dots \pi_V^{y_{vij}} \\
&= \left(\begin{array}{ccc} \pi_1^{y_{11j}} & \pi_2^{y_{21j}} & \dots & \pi_V^{y_{V1j}} \end{array} \right) \\
&\quad \times \left(\begin{array}{ccc} \pi_1^{y_{12j}} & \pi_2^{y_{22j}} & \dots & \pi_V^{y_{V2j}} \end{array} \right) \\
&\quad \times \dots \\
&\quad \times \left(\begin{array}{ccc} \pi_1^{y_{1ij}} & \pi_2^{y_{2ij}} & \dots & \pi_V^{y_{vij}} \end{array} \right) \\
&\quad \times \dots \\
&\quad \times \left(\begin{array}{ccc} \pi_1^{y_{1N_jj}} & \pi_2^{y_{2N_jj}} & \dots & \pi_V^{y_{VN_jj}} \end{array} \right) \\
&\quad \downarrow \qquad \qquad \qquad \downarrow \qquad \qquad \qquad \downarrow \\
&= \pi_1^{y_{11j}+y_{12j}+\dots+y_{1N_jj}} \\
&\qquad \qquad \qquad \pi_2^{y_{21j}+y_{22j}+\dots+y_{2N_jj}} \\
&\qquad \qquad \qquad \qquad \qquad \qquad \dots \pi_V^{y_{V1j}+y_{V2j}+\dots+y_{VN_jj}} \\
&= \pi_1^{N_{1j}} \\
&\qquad \qquad \qquad \pi_2^{N_{2j}} \\
&\qquad \qquad \qquad \qquad \qquad \qquad \dots \pi_V^{N_{Vj}} \\
&= \prod_{v=1}^V \pi_v^{N_{vj}}
\end{aligned}$$

最初 の等号：総積記号を分解して書き下した。

二番目の等号：縦方向に見方を変えて π ごとに見直した。

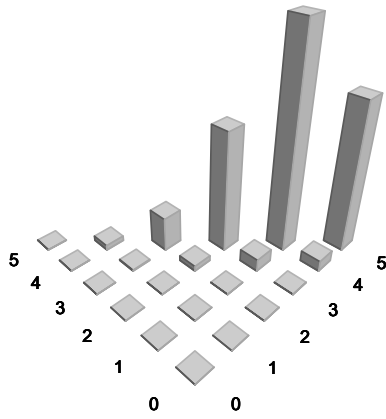
三番目の等号： j 番目の文書に出てくる単語 v の数を N_{vj} とおいた。ここで、 y_{vij} は、0 か 1 の二通りしかないことに注意！

四番目の等号：総積記号を使って、コンパクトに表現した。

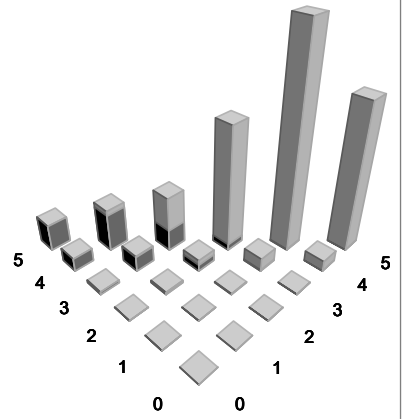


混合多項分布の例（二つの多項分布を混ぜた例）

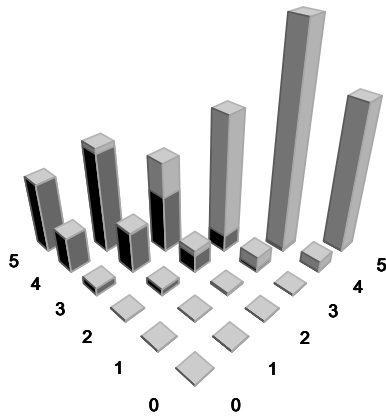
$\theta = 0$



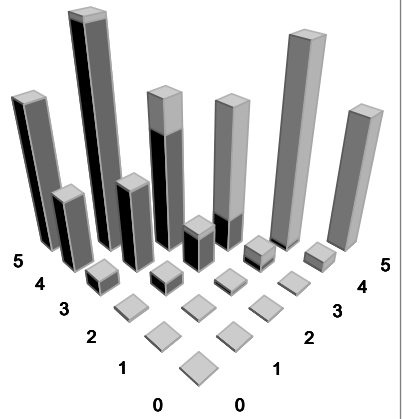
$\theta = 0.2$



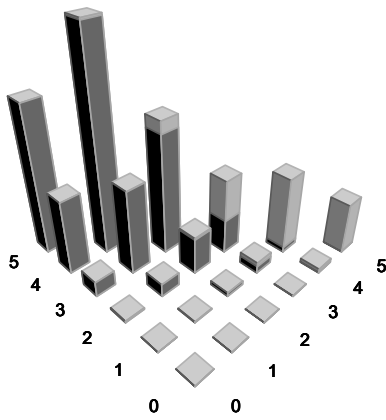
$\theta = 0.4$



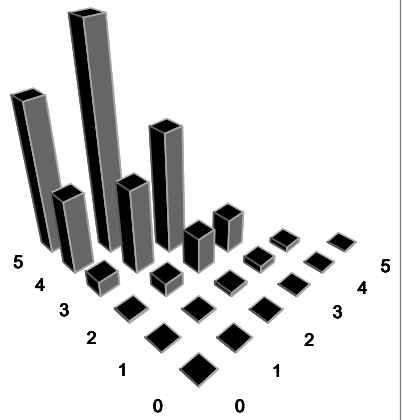
$\theta = 0.6$



$\theta = 0.8$



$\theta = 1$





得られた文書（データ）を生成する確率モデルを作ろう！

(1) ユニグラムモデル = 多項ロジスティック切片モデル

文書 j の i 番目のデータ（単語）が独立に同一のカテゴリカル分布から生まれると仮定。

$$y_{ij} \sim \text{Cat}(\phi)$$



Bag-of-Words

これは、「文」から「構造」を取り除いて作られた、「語」の集合。重複が許される多重集合である。

(1) [What I found] was [a huge dog [with a long tail]] .



(2) { a, a, dog, found, huge, I, long, tail, was, what, with }



「文書データ」を従属変数にする

Bag-of-words 表現によって、「文書データ」をベクトルと見なす。

$$(3) \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \\ \mathbf{y}_5 \\ \mathbf{y}_6 \\ \mathbf{y}_7 \\ \mathbf{y}_8 \\ \mathbf{y}_9 \\ \mathbf{y}_{10} \\ \mathbf{y}_{11} \end{pmatrix} = \begin{matrix} & \text{a} & \text{dog} & \text{find} & \text{huge} & \text{I} & \text{long} & \text{tail} & \text{be} & \text{what} & \text{with} \\ \begin{pmatrix} (0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0)^T \\ (0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0)^T \\ (1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0)^T \\ (1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0)^T \end{pmatrix} \end{matrix}$$

(データが一単語だけのとき)

i 番目の単語が産出される確率 :

$$\begin{aligned} p(\mathbf{y}_{ij}|\boldsymbol{\phi}) &= \phi_1^{y_{1ij}} \phi_2^{y_{2ij}} \dots \phi_V^{y_{Vij}} \\ &= \prod_{v=1}^V \phi_v^{y_{vij}} \end{aligned}$$

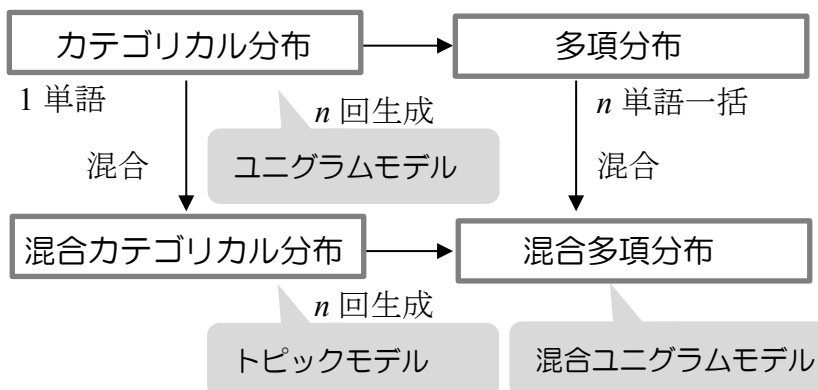
(データが n 単語あるとき)

1 番目から n 番目の単語が同時に産出される確率 :

$$\begin{aligned} p(\mathbf{y}_j|\boldsymbol{\phi}) &= p(\mathbf{y}_{1j}|\boldsymbol{\theta}, \boldsymbol{\phi}) p(\mathbf{y}_{2j}|\boldsymbol{\theta}, \boldsymbol{\phi}) \dots p(\mathbf{y}_{nj}|\boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= \left(\prod_{v=1}^V \phi_v^{y_{v1j}} \right) \left(\prod_{v=1}^V \phi_v^{y_{v2j}} \right) \dots \left(\prod_{v=1}^V \phi_v^{y_{vnj}} \right) \\ &= \phi_1^{y_{11j}} \phi_2^{y_{21j}} \dots \phi_V^{y_{V1j}} \\ &\quad \times \phi_1^{y_{12j}} \phi_2^{y_{22j}} \dots \phi_V^{y_{V2j}} \\ &\quad \times \phi_1^{y_{1nj}} \phi_2^{y_{2nj}} \dots \phi_V^{y_{Vnj}} \\ &= \phi_1^{y_{11j}+y_{12j}+\dots+y_{1nj}} \phi_2^{\sum_{i=1}^n y_{2ij}} \dots \phi_V^{\sum_{i=1}^n y_{Vij}} \\ &= \left(\prod_{v=1}^V \phi_v^{\sum_{i=1}^n y_{vij}} \right) \end{aligned}$$



見取り図：分布／モデル間の関係



※トピックモデルは混合したものを n 単語生成

※混合ユニグラムモデルでは n 単語生成させてから混合

(2) 混合ユニグラムモデル

これは、単語分布が文書のトピックに応じ変化するという想定した混合カテゴリカル分布を基にした統計モデル。

【ステップ1】トピックの生成 ← 混合比

j 番目の文書のトピック z_j は、1 から K の中から一つランダムに定まり、この分布をトピック分布と呼ぶ。

$$z_j \sim \text{Cat}(\boldsymbol{\theta})$$

$$p(z_{ij} = k | \boldsymbol{\theta}_j) = \theta_k$$

【ステップ2】単語の生成 ← 混合される前の分布

j 番目の文書の i 番目の単語が得られる確率はそのトピック z_j に応じたカテゴリカル分布（単語分布）に沿う。

$$\mathbf{y}_{ij} \sim \text{Cat}(\boldsymbol{\phi}_{z_j})$$

$$p(\mathbf{y}_{ij} | \boldsymbol{\theta}_j, \boldsymbol{\phi}, z_{ij} = k) = \phi_{1k}^{y_{1ij}} \phi_{2k}^{y_{2ij}} \dots \phi_{V_k}^{y_{vij}}$$

同一文書（トピック z_j ）下で n 回繰り返されると考える。

$$\mathbf{y}_j \sim \text{Multi}(\boldsymbol{\phi}_{z_j}, N_j)$$

$$p(\mathbf{y}_j | \boldsymbol{\theta}_j, \boldsymbol{\phi}, N_j, z_{ij} = k) = \phi_{1k}^{y_{1ij}} \phi_{2k}^{y_{2ij}} \dots \phi_{V_k}^{y_{vij}}$$

【ステップ3】確率密度関数

j 番目の文書が得られる確率密度式を、ステップ1とステップ2の積の和として求める。

$$p(\mathbf{y}_j | \boldsymbol{\theta}, \boldsymbol{\phi}, N_j) = \sum_{k=1}^K \theta_k \times p(\mathbf{y}_j | \boldsymbol{\phi}_k, N_j)$$



データの表記1 : 「文書 (作品)」と「文書集合 (作品集)」

文書が集まった「文書集合 (作品集)」が我々が手にするデータ。これを \mathbf{y} とし、その中の各「文書 (作品)」を $\mathbf{y}_1, \mathbf{y}_2, \dots$ のように表す。

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_J \end{pmatrix} = \begin{pmatrix} \text{羅生門} \\ \text{トロッコ} \\ \vdots \\ \text{蜘蛛の糸} \end{pmatrix}$$



データの表記2 : 「単語」と「文書 (作品)」

文書 \mathbf{y}_j は、 j 番目の文書の i 番目の単語を表すベクトルを考え、これを \mathbf{y}_{ij} とおく。

	老婆	男	門	線路	
$\mathbf{y}_1 = \text{羅生門} =$	$\begin{pmatrix} \mathbf{y}_{11} \\ \mathbf{y}_{21} \\ \mathbf{y}_{31} \\ \vdots \\ \mathbf{y}_{n_11} \end{pmatrix} =$	$\begin{pmatrix} (0 & 0 & 1 & \dots & 0)^T \\ (0 & 1 & 0 & \dots & 0)^T \\ (0 & 1 & 0 & \dots & 0)^T \\ \vdots \\ (1 & 0 & 0 & \dots & 0)^T \end{pmatrix}$			

		1	2	3	V	
		老婆	男	門	線路	
$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \vdots \\ \mathbf{y}_J \end{pmatrix} =$	$\begin{pmatrix} \begin{pmatrix} \mathbf{y}_{11} \\ \mathbf{y}_{21} \\ \mathbf{y}_{31} \\ \vdots \\ \mathbf{y}_{N_11} \end{pmatrix} \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \vdots \\ \mathbf{y}_J \end{pmatrix} =$	$\begin{pmatrix} \begin{pmatrix} (0 & 0 & 1 & \dots & 0)^T \\ (0 & 1 & 0 & \dots & 0)^T \\ (0 & 1 & 0 & \dots & 0)^T \\ \vdots \\ (1 & 0 & 0 & \dots & 0)^T \end{pmatrix} \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \vdots \\ \mathbf{y}_J \end{pmatrix}$				

(3) トピックモデル

これは、単語の生成過程のモデルとして、「文書 j 」ごとに異なる混ざり方をした単語分布を想定する統計モデル。

【ステップ1】 トピックの生成 ← 混合比

j 番目の文書の単語 i のトピック z_{ij} は、文書ごとに定められるトピック分布からランダムに決まる。

$$z_{ij} \sim \text{Cat}(\boldsymbol{\theta}_j)$$

$$p(z_{ij} = k | \boldsymbol{\theta}_j) = \theta_{jk}$$

【ステップ2】 単語の生成 ← 混合された後の分布

j 番目の文書に含まれる i 番目の単語が $\mathbf{y}_{ij} = (y_{1ij} \ y_{2ij} \ \dots \ y_{vij})^T$ になる確率はトピック z_{ij} の単語分布が規定。

$$\mathbf{y}_{ij} \sim \text{Cat}(\boldsymbol{\phi}_{z_{ij}})$$

$$\begin{aligned} p(\mathbf{y}_{ij} | \boldsymbol{\theta}_j, \boldsymbol{\phi}) &= \sum_{k=1}^K p(z_{ij} = k | \boldsymbol{\theta}_j) p(\mathbf{y}_{ij} | \boldsymbol{\phi}_k) \\ &= \sum_{k=1}^K \theta_k (\phi_{1k}^{y_{1ij}} \phi_{2k}^{y_{2ij}} \dots \phi_{vk}^{y_{vij}}) \end{aligned}$$

【ステップ3】 確率密度関数

j 番目の文書が得られる確率は、同一文書 (トピック z_j) 下でステップ2が n 回繰り返されたものとする。

$$\begin{aligned} p(\mathbf{y}_j | \boldsymbol{\theta}_j, \boldsymbol{\phi}, N_j) &= \left[\sum_{k=1}^K \theta_k (\phi_{1k}^{y_{11j}} \phi_{2k}^{y_{21j}} \dots \phi_{vk}^{y_{v1j}}) \right] \\ &\quad \times \left[\sum_{k=1}^K \theta_k (\phi_{1k}^{y_{12j}} \phi_{2k}^{y_{22j}} \dots \phi_{vk}^{y_{v2j}}) \right] \\ &\quad \times \dots \times \left[\sum_{k=1}^K \theta_k (\phi_{1k}^{y_{1N_j j}} \phi_{2k}^{y_{2N_j j}} \dots \phi_{vk}^{y_{vN_j j}}) \right] \\ &= \prod_{i=1}^{N_j} \left[\sum_{k=1}^K \theta_k (\phi_{1k}^{y_{1ij}} \phi_{2k}^{y_{2ij}} \dots \phi_{vk}^{y_{vij}}) \right] \end{aligned}$$



混合ユニグラムモデルとトピックモデルの比較

混合ユニグラムモデル

- ① トピック z_j 文書 j ごとに決定
- ② トピックが生じる比率 θ 文書 j には無関係

トピックモデル

- ① トピック z_{ij} 文書 j の各単語 i ごとに決定
- ② トピックが生じる比率 θ_j 文書 j ごとに決定



混合ユニグラムモデルの基本的な考え方

(例) 世の作品に「恋愛」と「友情」という2つのトピックがあると想定する(3つ目、例えば「生と死」を加えても話は同じ)。

「 j 番目の文書の i 番目の単語が『男』である確率」

= 「① j 番目の文書のトピックが『恋愛』で、② この『恋愛』の単語分布から『男』が出る確率

+

① j 番目の文書のトピックが『友情』で、② この『友情』の単語分布から『男』が出る確率

= $\theta_{\text{恋愛}} \times p(y_{ij}|z_j = \text{恋愛}) + \theta_{\text{友情}} \times p(y_{ij}|z_j = \text{友情})$

↑ ↑ ↑ ↑
混合比 確率密度 + 混合比 確率密度



トピックモデルの基本的な考え方

「 j 番目の文書の i 番目の単語が『男』である確率」

= 「① j 番目の文書の i 番目の単語のトピックが『恋愛』で、
 ② この『恋愛』の単語分布から『男』が出る確率

+

① j 番目の文書の i 番目の単語のトピックが『友情』で、
 ② この『友情』の単語分布から『男』が出る確率

= $\theta_{\text{恋愛}} \times p(y_{ij}|z_{ij} = \text{恋愛}) + \theta_{\text{友情}} \times p(y_{ij}|z_{ij} = \text{友情})$

↑ ↑ ↑ ↑
混合比 確率密度 + 混合比 確率密度

例題**TASA コーパスの分析**

Steyvers and Griffiths (2007)で紹介されている TASA コーパス (Touchstone Applied Science Associates によって作成された教育的内容の 37,000 個の文書コーパス) のトピックモデルによる解析結果 (※300 個のトピックを想定している) を見て、77, 82, 166 のトピックがどのようなカテゴリーを形成しているか、その潜在的意味の候補を推測してみよう。なお、問題の作成上、修正を施した箇所がある。

Document #29795

Bix Beiderbecke, at age⁰⁶⁰ fifteen²⁰⁷, sat¹⁷⁴ on the slope⁰⁷¹ of a bluff⁰⁵⁵ overlooking⁰²⁷ the Mississippi¹³⁷ river¹³⁷. He was listening⁰⁷⁷ to music⁰⁷⁷ coming⁰⁰⁹ from a passing⁰⁴³ riverboat. The music⁰⁷⁷ had already captured⁰⁰⁶ his heart¹⁵⁷ as well as his ear¹¹⁹. It was jazz⁰⁷⁷. Bix Beiderbecke had already had music⁰⁷⁷ lessons⁰⁷⁷. He showed⁰⁰² promise¹³⁴ on the piano⁰⁷⁷, and his parents⁰³⁵ hoped²⁶⁸ he might consider¹¹⁸ becoming a concert⁰⁷⁷ pianist⁰⁷⁷. But Bix was interested²⁶⁸ in another kind⁰⁵⁰ of music⁰⁷⁷. He wanted²⁶⁸ to play⁰⁷⁷ the cornet. And he wanted²⁶⁸ to play⁰⁷⁷ jazz⁰⁷⁷.

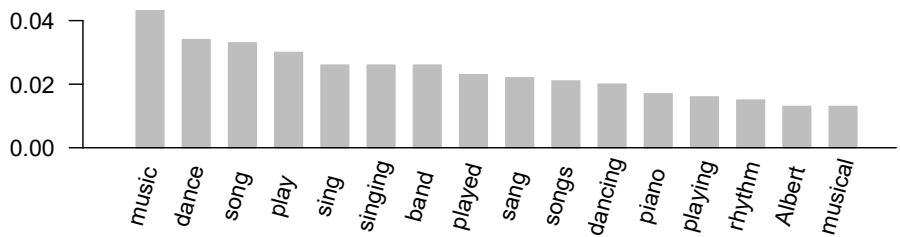
Document #1883

There is a simple⁰⁵⁰ reason¹⁰⁶ why there are so few periods⁰⁷⁸ of really great theater⁰⁸² in our whole western⁰⁴⁶ world. Too many things³⁰⁰ have to come right at the very same time. The dramatists must have the right actors⁰⁸², the actors⁰⁸² must have the right playhouses, the playhouses must have the right audiences⁰⁸². We must remember²⁸⁸ that plays⁰⁸² exist¹⁴³ to be performed⁰⁷⁷, not merely⁰⁵⁰ to be read²⁵⁴. (even when you read²⁵⁴ a play⁰⁸² to yourself, try²⁸⁸ to perform⁰⁶² it, to put¹⁷⁴ it on a stage⁰⁷⁸, as you go along.)

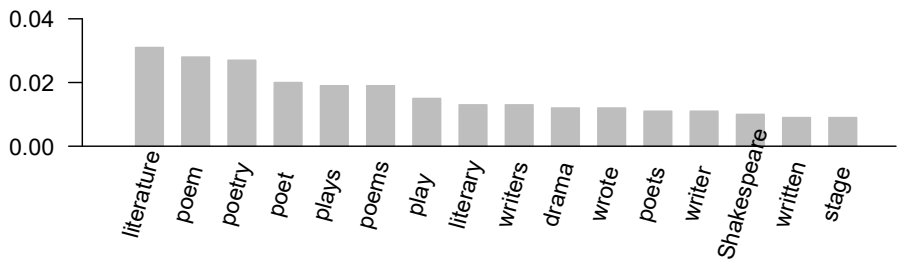
Document #21359

Jim has a game¹⁶⁶ book²⁵⁴. Jim reads²⁵⁴ the book²⁵⁴. Jim sees⁰⁸¹ a game¹⁶⁶ for one. Jim plays¹⁶⁶ the game¹⁶⁶. Jim likes⁰⁸¹ the game¹⁶⁶ for one. The game¹⁶⁶ book²⁵⁴ helps⁰⁸¹ Jim. Don comes⁰⁴⁰ into the house⁰³⁸. Don and Jim read²⁵⁴ the game¹⁶⁶ book²⁵⁴. The boys⁰²⁰ see a game¹⁶⁶ for two. The two boys⁰²⁰ play¹⁶⁶ the game¹⁶⁶. The boys⁰²⁰ play¹⁶⁶ the game¹⁶⁶ for two. The boys⁰²⁰ like the game¹⁶⁶.

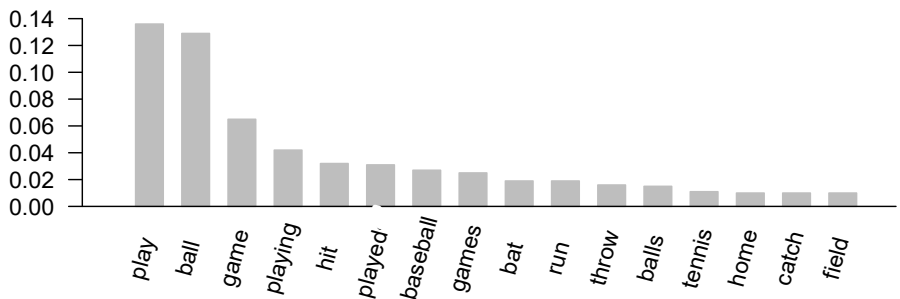
Topic 77



Topic 82



Topic 166



Steyvers, Mark & Tom Griffiths. (2007) Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*, ed. by Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, Walter Kintsch, 439-460. New York: Psychology Press.