ノンパラメトリックベイズ入門
～無限 HMM と教師なし単語分割～

Graham Neubig

２０１０年１１月２９日

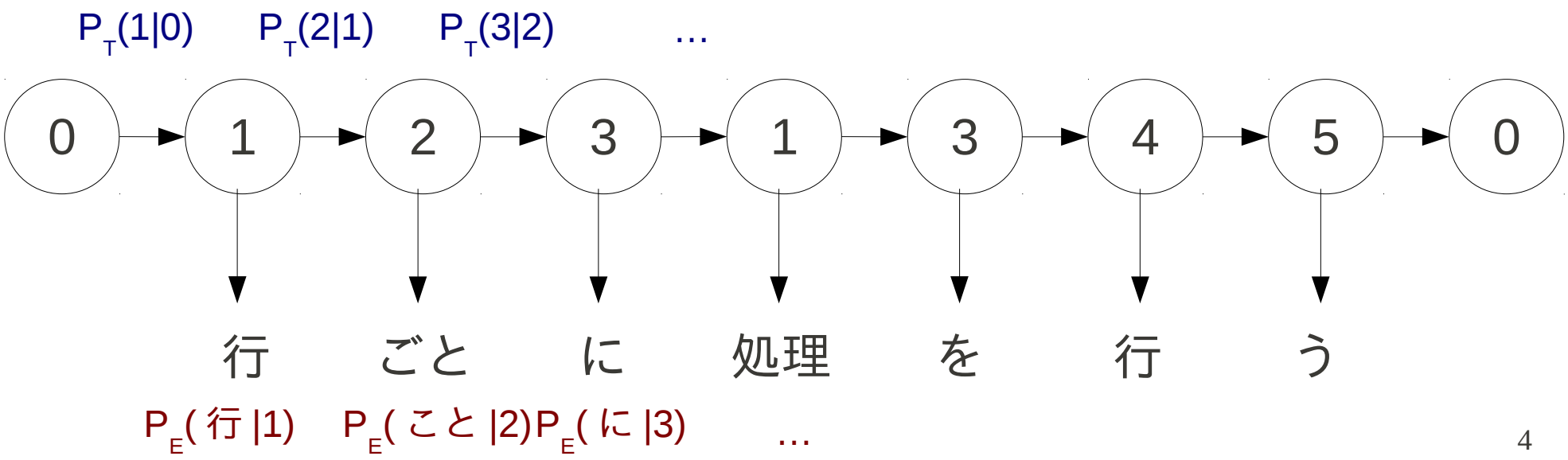
発表の概要

- 前回はディリクレ過程、基底測度、ギブスサンプリングなどを紹介した
- しかし、モデルはまだパラメトリック
  - パラメータの数は限られていた
- 今回は真のノンパラメトリックベイズ
  - 無限 HMM （ HMM の状態数は無限）
  - 教師なし単語分割（可能な単語の数は無限）

有限 HMM から無限 HMM へ

HMM のモデル（復習）

- 品詞は隠れている状態
  - 状態の遷移確率は $P_T(y_i|y_{i-1})=\theta_{T,y_i,y_{i-1}}$
- 各状態から単語を生成する
  - 状態が与えられた場合の生成確率は $P_E(x_i|y_i)=\theta_{E,y_i,x_i}$



ディリクレ過程

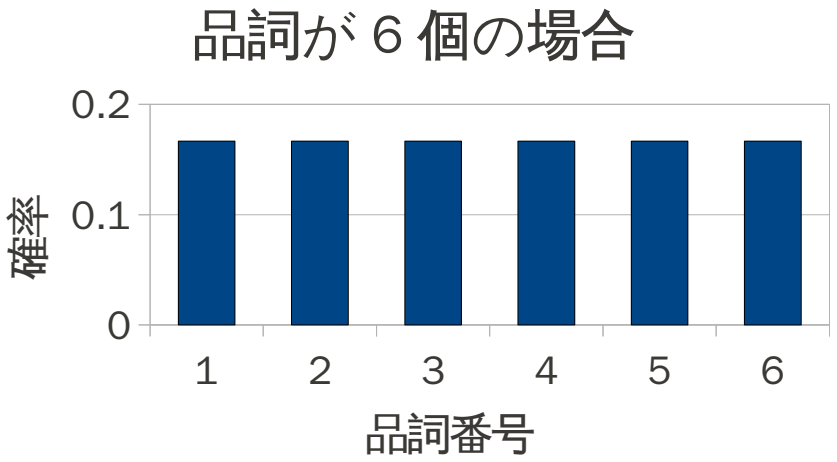
- 前回の実装で利用した遷移確率を思い出してみよう

$$P(y_i|y_{i-1}) = \frac{c(y_{i-1}y_i) + \alpha * P_{base}(y_i)}{c(y_{i-1}) + \alpha}$$

- ディリクレ過程を利用すると
  - 最尤推定で利用する頻度
  - 加算スムージングのような項
- ここで鍵を握るのは基底測度 P_{base}

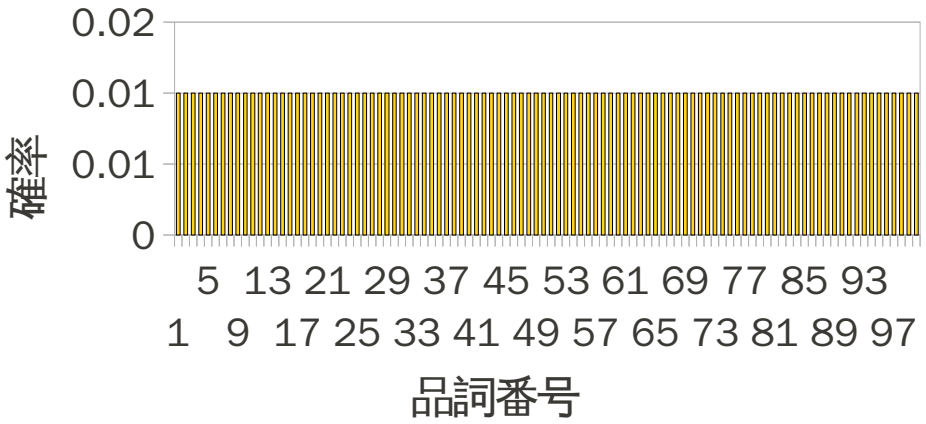
基底測度の次元数

- 実装の課題で一様分布の基底測度を利用した

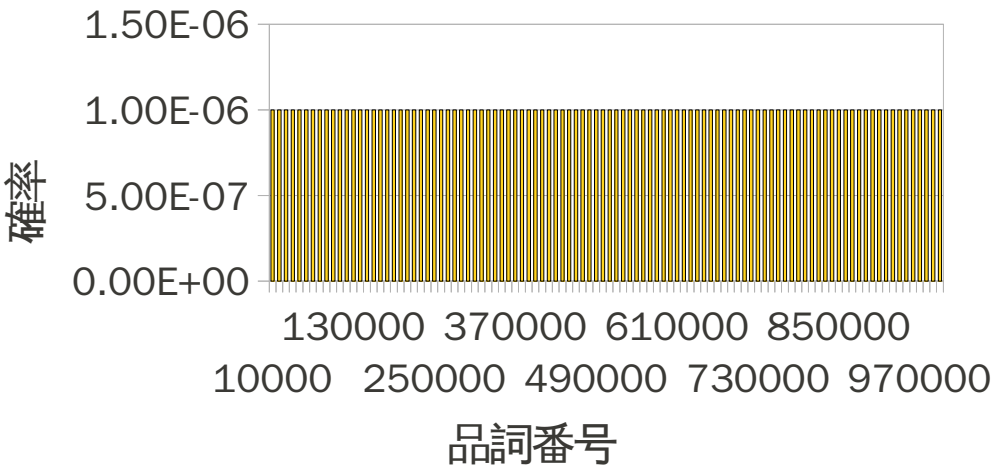


さらに伸ばすと

品詞が 100 個の場合



品詞が 100 万個の場合



- 品詞の数を極大まで増やしていくと
  - それぞれの品詞の P_{base} がゼロに近づく
  - しかし、 P_{base} 全体から品詞を生成する確率はそのまま

$$P(y_i|y_{i-1}) = \frac{c(y_{i-1}y_i) + \alpha * P_{base}(y_i)}{c(y_{i-1}) + \alpha}$$

N= 品詞の数

$$\lim_{N\to\infty} \sum_{i=1}^{N} \frac{1}{N} = 1$$

有限 HMM と無限 HMM

- 有限 HMM
  存在する品詞 y_i を (y_{i-1} の後に) 生成する確率

$$P(y_i|y_{i-1}) = \frac{c(y_{i-1}y_i) + \alpha * P_{base}(y_i)}{c(y_{i-1}) + \alpha}$$

- 無限 HMM
  存在する品詞 y_i を
  (y_{i-1} の後に) 生成する確率

$$P(y_i|y_{i-1}) = \frac{c(y_{i-1}y_i)}{c(y_{i-1}) + \alpha}$$

  新しい品詞を
  (y_{i-1} の後に) 生成する確率

$$P(y_i|y_{i-1}) = \frac{\alpha}{c(y_{i-1}) + \alpha}$$

例えば

- $c(y_{i-1}=1\ y_i=1)=1$ $c(y_{i-1}=1\ y_i=2)=1$ としよう

可能な品詞が 2 個 (y_1,y_2) の場合

$$P(y_i=1|y_{i-1}=1)=\frac{1+\alpha*1/2}{2+\alpha} \quad P(y_i=2|y_{i-1}=1)=\frac{1+\alpha*1/2}{2+\alpha}$$

$$P(y_i=1,2\text{以外}|y_{i-1}=1)=\frac{\alpha*0}{2+\alpha}$$

可能な品詞が 20 個 (y_1,y_{20}) の場合

$$P(y_i=1|y_{i-1}=1)=\frac{1+\alpha*1/20}{2+\alpha} \quad P(y_i=2|y_{i-1}=1)=\frac{1+\alpha*1/20}{2+\alpha}$$

$$P(y_i=1,2\text{以外}|y_{i-1}=1)=\frac{\alpha*18/20}{2+\alpha}$$

可能な品詞が∞個 (y_1,y_∞) の場合

$$P(y_i=1|y_{i-1}=1)=\frac{1+\alpha*1/\infty}{2+\alpha} \quad P(y_i=2|y_{i-1}=1)=\frac{1+\alpha*1/\infty}{2+\alpha}$$

$$P(y_i=1,2\text{以外}|y_{i-1}=1)=\frac{\alpha*1}{2+\alpha}$$

無限とは？

- ここではなぜ「無限の可能なタグ」にするか
- 理論：タグが多くても、まだまだ見たことがない珍しい言語現象はあるかもしれない
  - 「待ったをかける」
  - 「紫のかゞやく花と日の光思ひあはざることわりもなし」
- 実用：コーパスが大きくなればなるほどたくさんのクラスが欲しい

サンプリングアルゴリズム

SampleTag(y_i)

$c(y_{i-1}\ y_i)$--; $c(y_i\ y_{i+1})$--; $c(y_i \rightarrow x_i)$--

今のタグのカウントを削除する

for each *tag* in S (品詞の集合)

$p[tag]=P_E(tag|y_{i-1})*P_E(y_{i+1}|tag)*P_T(x_i|tag)$

存在するタグの確率を計算（ディリクレ過程の式で，p.8）

$p[|S|+1]=P_E(new|y_{i-1})*P_E(y_{i+1}|new)*P_T(x_i|new)$

新しいタグの確率を計算

y_i = **SampleOne**(p)

y_i の値を選ぶ

$c(y_{i-1}\ y_i)$++; $c(y_i\ y_{i+1})$++; $c(y_i \rightarrow x_i)$++

そのタグを追加する

実装のいろいろ

- 普通は 0 頻度のクラスは残る→クラス数が上がる一方
  - 新しいクラスを作る時に 0 頻度のクラス番号を使い回す

$c(y_1)=5\ c(y_2)=0\ c(y_3)=1$ →
単純： $c(y_1)=5\ c(y_2)=0\ c(y_3)=1\ c(y_4)=1$
賢い： $c(y_1)=5\ c(y_2)=1\ c(y_3)=1$

  - $c(y_1)=0$ になると、 y_1 が復活する確率が 0 になる
- このモデルだと新しい品詞に弱い
  - 新しい品詞は 1 つの品詞の後にしか来ない
  - 階層モデルで解決

遷移確率 → $P_T(y_i|y_{i-1})=DP(\alpha,P_T(y_i))$

品詞確率 → $P_T(y_i)=DP(\alpha,P_{base}(y_i))$

教師なし単語分割のモデル

言語モデルを使った教師あり単語分割

- 日本語や中国語などは明白な区切りがなくても単語という概念はある→単語分割は必要
- 言語モデルを使った単語分割はできる
  - 分割したい文字列 **x** からなるすべての単語列 **w** は分割候補
  - 候補の確率は言語モデル確率に比例
  - 確率の最も高い候補を正しい分割

1-gram 言語モデル

- 一番簡単な言語モデルは 1-gram
  - 文のそれぞれの単語は独立に生成

$$P(\boldsymbol{w}) = \prod_{w_i \in \boldsymbol{w}} P(w_i)$$

- 1-gram 確率にディリクレ過程の事前確率をかける

$$P(w_i) = \frac{c(w_i) + \alpha P_{base}(w_i)}{\sum_{\tilde{i}} c(w_{\tilde{i}}) + \alpha}$$

- 2-gram や 3-gram を利用すると精度が上がる (上級編)

単語分割の基底測度

- 単語に対する基底測度は単語になり得る全ての文字列に確率を与えなければならない
- よく使われているモデルは：
  - まず単語の長さ J を確率的に選ぶ

$$P_{len}(J)$$

  - それぞれの J 文字を文字の確率分布から独立に選ぶ

$$\prod_{j=1}^{J} P_{char}(x_j)$$

  - 組み合わせると、どの単語でも少しの確率が振られる

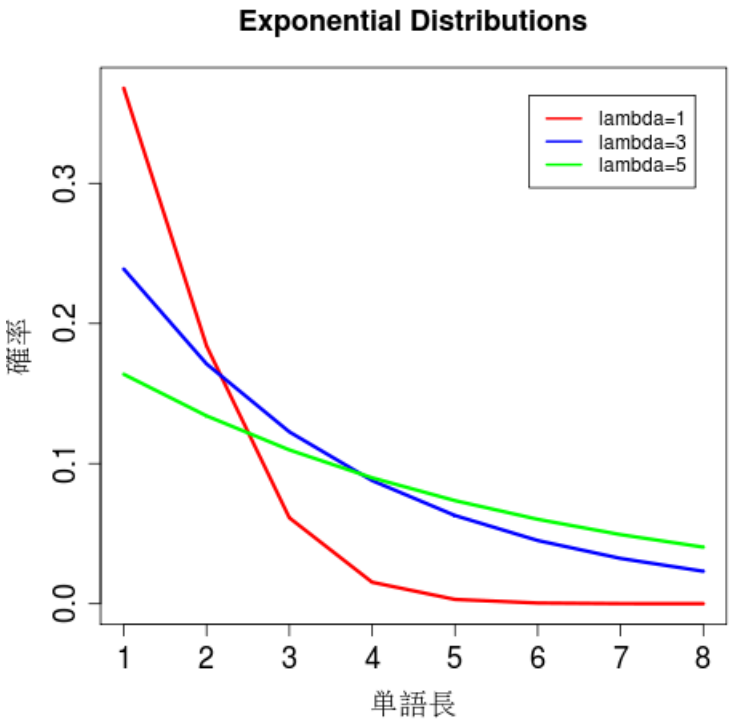
$w_i = x_1 \ldots x_J$ $\qquad$ $P_{base}(w_i) = P_{len}(J) \prod_{j=1}^{J} P_{char}(x_j)$

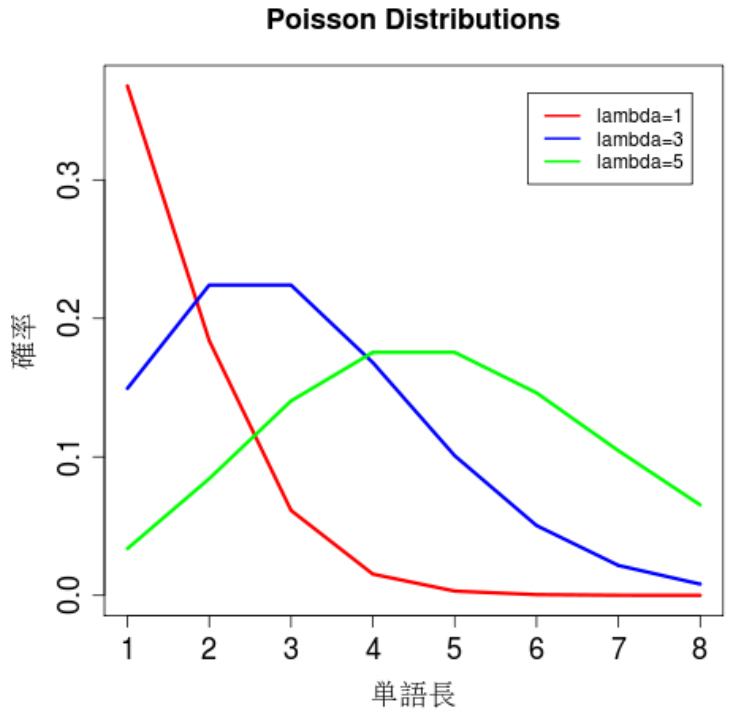
長さの確率

- 長さの確率分布として指数分布とポアソン分布

指数分布： $P_{len}(J)=\frac{1}{\lambda}e^{-\frac{J}{\lambda}}$

ポアソン分布： $P_{len}(J)=\lambda^{J}\frac{e^{(-\lambda)}}{J!}$





- 両方は「平均単語長λ」のパラメータがある

文字 1-gram の分布

- 文字 1-gram は予め最尤推定で計算しても良い

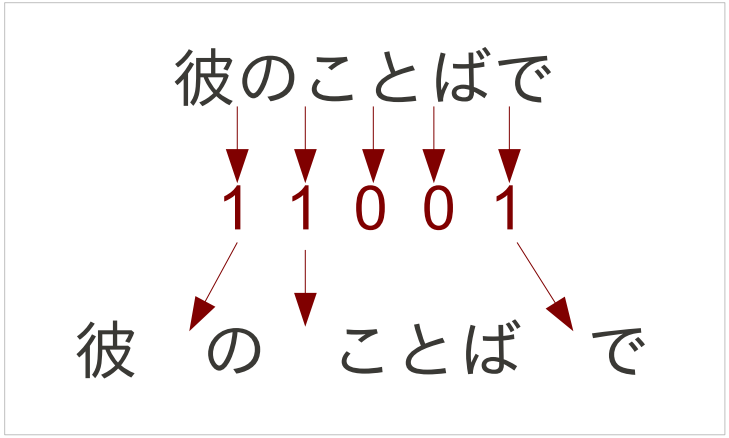
$$P_{char}(x_j) = \frac{c(x_j)}{\sum_{\tilde{j}} c(x_{\tilde{j}})}$$

- 未知の文字が扱えなくなるが、とりあえず学習コーパスを分割したいので、気にしない
- 文字 n-gram を利用したり、カタカナ / 漢字 / 数字の違いを利用するモデルもある（上級編、持橋 + ）

教師なし単語分割のサンプリング

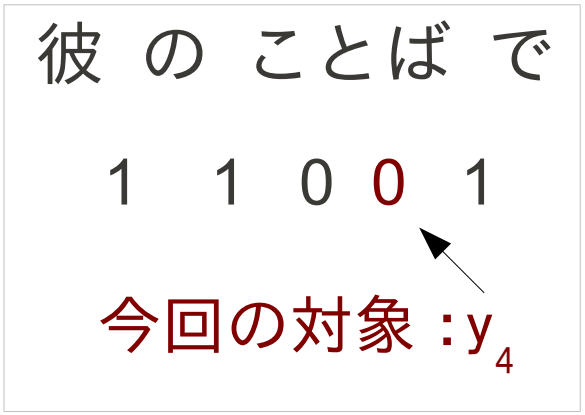
単語分割の表現方法

- 分割された文を表す方法として
  - 文字列 **x** ： $x_1\ x_2\ x_3\ x_4\ x_5\ x_6$ 彼のことばで （観測変数）
  - 単語境界タグ列 **y** ： $y_1\ y_2\ y_3\ y_4\ y_5$ 1 1 0 0 1 （非観測変数）
- y_i が 1 であれば、 x_i と x_{i+1} の間に単語境界がある



サンプリングの対象

- サンプリングする変数は 1 個の単語境界タグ y_i の値



- 1-gram モデルで影響されるのはその単語境界に隣接する単語だけ（「こと」と「ば」または「ことば」）

P(彼) * P(の) *	P(ことば)	* P(で)
P(彼) * P(の) *	P(こと) * P(ば)	* P(で)

サンプリングの過程

- まず、単語境界に隣接する単語の頻度から１を引く

c(ことば)--　or　c(こと)--　c(ば)--

- コーパスの総単語数を記録（保持）しておく

$$W = \sum_{\tilde{i}} c(w_{\tilde{i}})$$

サンプリングの過程

- 単語境界タグが 0 の場合の確率：１単語

$$P(y_i=0) \propto P(\text{ことば}) = \frac{c(\text{ことば}) + \alpha P_{base}(\text{ことば})}{W + \alpha}$$

- 単語境界タグが 1 の場合の確率：２単語

$$P(y_i=1) \propto P(\text{こと}) P(\text{ば})$$

$$P(\text{こと}) = \frac{c(\text{こと}) + \alpha P_{base}(\text{こと})}{W + \alpha}$$

$$P(\text{ば}) = \frac{c(\text{ば}) + \alpha P_{base}(\text{ば})}{W + 1 + \alpha}$$

「こと」が追加されるため W+1

なぜ 1 を追加

- ディリクレ過程で単語が生成されるたびに確率が変わる

これ $P(\text{これ}) = \frac{c(\text{これ}) + \alpha P_{base}(\text{これ})}{W + \alpha} = \frac{0 + \alpha P_{base}(\text{これ})}{0 + \alpha}$

c(これ)++ W++

は $P(\text{は}) = \frac{c(\text{は}) + \alpha P_{base}(\text{は})}{W + \alpha} = \frac{0 + \alpha P_{base}(\text{は})}{1 + \alpha}$

c(は)++ W++

これ $P(\text{これ}) = \frac{c(\text{これ}) + \alpha P_{base}(\text{これ})}{W + \alpha} = \frac{1 + \alpha P_{base}(\text{これ})}{2 + \alpha}$

c(これ)++ W++

で $P(\text{で}) = \frac{c(\text{で}) + \alpha P_{base}(\text{で})}{W + \alpha} = \frac{0 + \alpha P_{base}(\text{で})}{3 + \alpha}$

c(で)++ W++

サンプリングの過程

- P(y_i=0) と P(y_i=1) に従って、y_iの新しい値をサンプリングする

$$P(y_i=0)=\frac{P(\text{ことば})}{P(\text{ことば})+P(\text{こと})P(\text{ば})}$$

- 「こと」と「ば」、または「ことば」の頻度に 1 を足す

```
c(ことば)++   or   c(こと)++  c(ば)++
```

実装のいろいろ

- 初期化はどうする？
  - 全部の文を１つの単語と初期化する
  - ランダムに初期化する
- 前向き後ろ向きアルゴリズムで１文ごとにサンプリングすることもできる (上級編：持橋 + ）
- タイプ（同じ単語になる境界）ごとにサンプリングもできる（上級編： Liang+ “Type based MCMC” ）

課題

課題

- 教師なし単語分割を実装すること
- 前回の課題で利用したデータで
- α の設定は実験的に決める
- 単語分割精度で評価（スクリプトは送っておきます）
  - 単語列のマッチングをかけて、 F 値で評価

さらに詳しく知りたい人に

- ノンパラメトリックベイズの紹介
  Jordan “Dirichlet Processes, Chinese Restaurant Processes, and all that.” (ビデオ)
  持橋　“最近のベイズ理論の進展と応用 (III) ノンパラメトリックベイズ”
- ノンパラメトリックベイズを使った言語モデル
  Teh “A Bayesian Interpretation of Interpolated Kneser Ney”
- ノンパラメトリックベイズを使った単語分割
  Goldwater+ “A Bayesian framework for word segmentation: exploring the effects of context”
  持橋 + “ ベイズ階層言語モデルによる教師なし形態素解析”