Query Rewrite Based On Document Model

工虫

Recall Based On Distribution

- current recall method:
  - q=ABCD
  - A && B && C && D
- denote: q={T_i} , d={T_i}
- assumption: the quantity of recall result is proportional to the similarity between two distributions of Q and D
- problem: P(D) = ?

Some Distributions

- Binominal distribution
- voting/coin toss problem:

$$Bin(x \mid n, \theta) = \begin{pmatrix} n \\ k \end{pmatrix} \theta^k (1-\theta)^{n-k} = \begin{pmatrix} n \\ k \end{pmatrix} P(x \mid \theta)$$

- MLE: $\max(\theta^k (1-\theta^{n-k}))$
- Bayesian estimition:

$$P(\theta \mid x) = \frac{P(x \mid \theta) P(\theta)}{P(x)} \qquad P(x) = \int_0^1 P(x \mid \theta) P(\theta) d\theta$$

Some Distributions

- Beta-Binomial distribution

$$\theta \sim Beta(a,b), P(\theta) = Beta(\theta \mid a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

$$P(\theta \mid x) = \frac{P(x \mid \theta)P(\theta)}{\int_0^1 P(x \mid \theta)P(\theta)\,d\theta}$$

$$= \frac{\theta^k(1-\theta)^{n-k}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 \theta^k(1-\theta)^{n-k}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\,d\theta}$$

$$= \frac{\theta^k(1-\theta)^{n-k}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 \theta^k(1-\theta)^{n-k}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\,d\theta}$$

$$= \frac{\theta^{k+a-1}(1-\theta)^{n-k+b-1}}{\int_0^1 \theta^{k+a-1}(1-\theta)^{n-k+a-1}\,d\theta}$$

$$= Beta(x \mid k+a-1, n-k+b-1)$$

Some Distributions

- example: coin toss/voting/ctr

Some Distributions

- Multinomial distribution & Dirichlet-Multinomial distribution
  - Multinomial: $Mu(x \mid n, \theta) = \binom{n}{x_1, x_2, \ldots} \prod_{j=1}^{k} \theta^{x_j}$
  - Dirichlet-Multinomial:

$$P(\theta) = Dir(\theta \mid \alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^{k} \theta^{\alpha_k - 1}$$

$$P(\theta \mid x) \propto P(x \mid \theta) P(\theta) = \frac{1}{B(\alpha)} \prod_{j=1}^{k} \theta^{x_j} \theta^{\alpha_k - 1} = Dir(\theta \mid \alpha_1 + x_1, \alpha_2 + x_2, \ldots)$$

$$s.t.$$

$$\sum \theta = 1$$

$$x \in S_k$$

Dirichlet Compound Model

- DCM:

$$P(x) = \int P(x \mid \theta) P(\theta) = \int Mu(x \mid n, \theta) Dir(\theta \mid \alpha) d\theta = \frac{n!}{\prod_{w=1}^{W} x_w !} \frac{B(x_w + \alpha_w)}{B(\alpha)}$$

- Advantage vs other document models:
  - considering the occurrence of a word (bustiness)

Approximation for DCM

- DCM:

$$B(\alpha) = \frac{\prod_{w=1}^{W} \Gamma(\alpha_w)}{\Gamma(\sum_w \alpha_w)}$$

$$s = \sum_w \alpha_w$$

$$P(x) = \frac{n!}{\prod_{w=1}^{W} x_w!} \frac{B(x_w + \alpha_w)}{B(\alpha)} = \frac{n!}{\prod_{w=1}^{W} x_w!} \frac{\Gamma(s)}{\Gamma(s+n)} \prod_{w=1}^{W} \frac{\Gamma(x_w + \alpha_w)}{\Gamma(\alpha_w)}$$

- approximation:

$$\lim_{\alpha \to 0} \frac{\Gamma(x+\alpha)}{\Gamma(\alpha)} - \Gamma(x)\alpha = 0$$

$$P(x) \approx Q(x) = \frac{n!}{\prod_{w=1}^{W} x_w!} \frac{\Gamma(s)}{\Gamma(s+n)} \prod_{w=1}^{W} \Gamma(x_w + \alpha_w)\alpha_w = \frac{n!}{\prod_{w=1}^{W} x_w!} \frac{\Gamma(s)}{\Gamma(s+n)} \prod_{w=1}^{W} \beta_w$$

Solving

- exp-form: $Q(x) = (\prod_{w=1}^{W} x_w^{-1}) n! \frac{\Gamma(s)}{\Gamma(s+n)} \exp[\sum_{w=1}^{W} I(x_w > 0) \log \beta_w]$
- MLE: $L(x) = \log n! + \log \Gamma(x) - \log \Gamma(s+n) + \sum_{w=1}^{W} (\log \beta_w - \log x_w)$

$$\frac{dL}{d\beta} = D\Psi(s) - \sum_{d=1}^{D} \Psi(s+n_d) + \sum_{d=1}^{D} I(x_{dw} > 0) \frac{1}{\beta_w}$$

$$\beta_w = \frac{\sum_{d=1}^{D} I(x_{dw} > 0)}{\sum_{d=1}^{D} \Psi(s+n_d) - D\Psi(s)}$$

$$s = \frac{\sum_{w=1}^{W} \sum_{d=1}^{D} I(x_{dw} > 0)}{\sum_{d=1}^{D} \Psi(s+n_d) - D\Psi(s)}$$

Solving

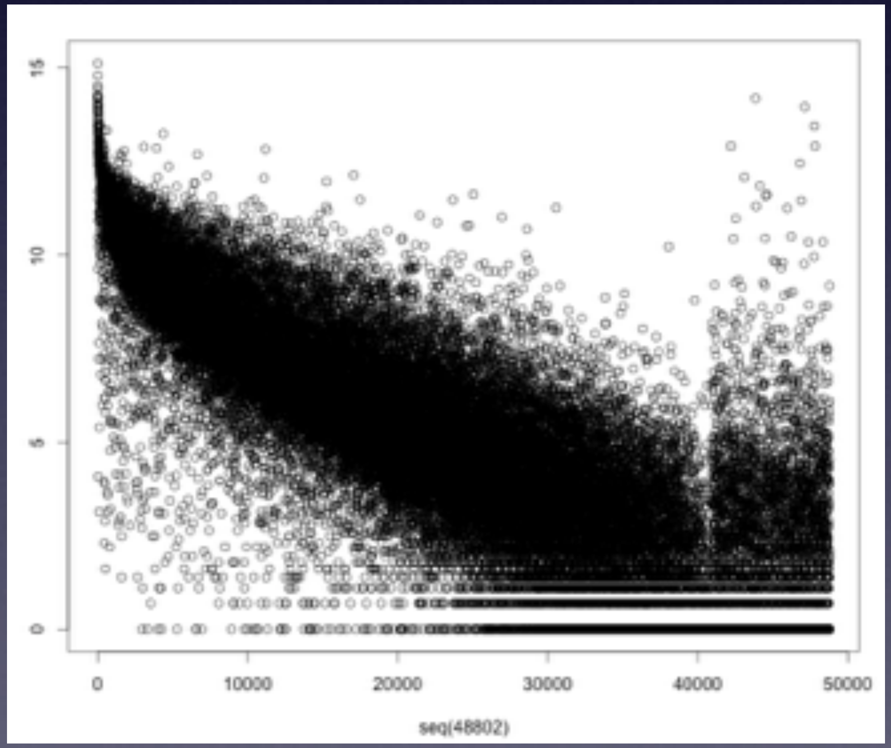
- Newton method

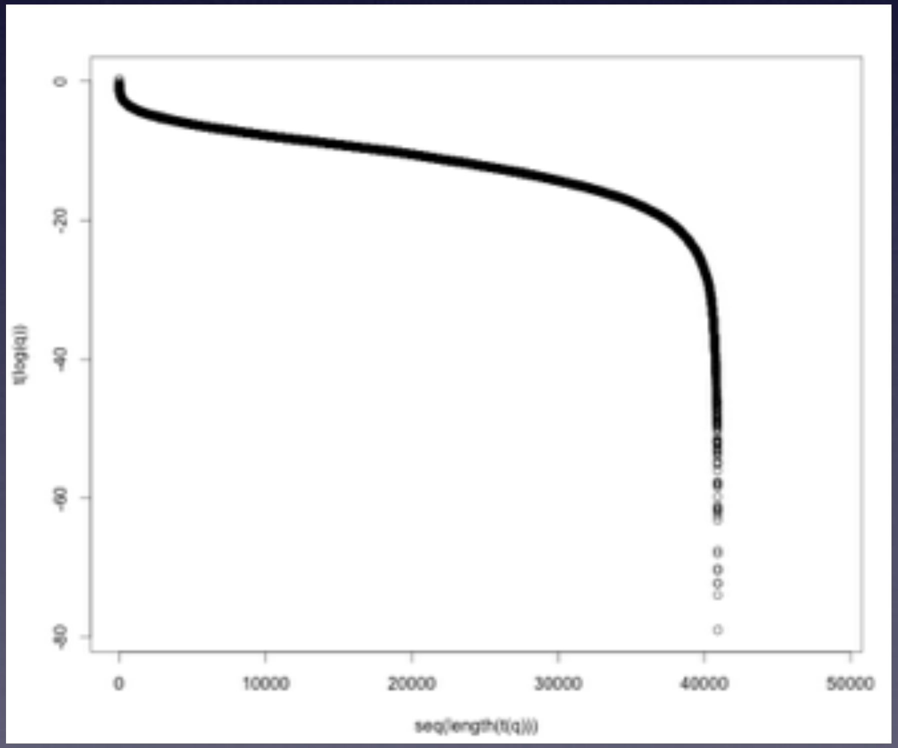
$$s = \frac{\sum_{w=1}^{W}\sum_{d=1}^{D} I(x_{dw} > 0)}{\sum_{d=1}^{D}\Psi(s+n_d) - D\Psi(s)} = \frac{C}{\sum_{d=1}^{D}\Psi(s+n_d) - D\Psi(s)}$$

$$F(s) = s(\sum_{d=1}^{D}\Psi(s+n_d) - D\Psi(s)) - C = 0$$

Experiments

- recall amount of query VS P(q)





Disadvantage VS Advantage

- disadvantage:
  - computational complexity (did not develop a parallel version)
  - T(x) is not computational when x > 200, that leads a bit error of the last result
- advantage:
  - consider of the word bustiness
  - easy to develop a version of online training
  - easy to develop a mixture model
  - fit for BOW and is much easier to use than LDA

END
THANKS