# Theory of TopWORDS

### Yu Huang 1

<sup>1</sup>yuhuang-cst@foxmail.com

#### 1. Notation

• Characters:  $A = \{a_1, a_2, ..., a_p\}$ 

• Vocabulary:  $D = \{w_1, w_2, ..., w_N\}, w = a_{i_1}a_{i_2}...a_{i_l}$ 

• Word probability:  $\boldsymbol{\theta} = (\theta_1, ..., \theta_N); \sum_{i=1}^N \theta_i = 1$ 

• K-word (segmented) sentence:  $S = w_{i_1}, w_{i_2}, ..., w_{i_K}$ 

• Unsegmented text T

• Set of all segmented sentences of T:  $C_T$ 

• Max word length:  $\tau_L$ 

• Corpus:  $T = \{T_1, ..., T_n\}$ 

• Number of texts in corpus: n

• Number of words in vocabulary: N

#### 2. Methods

Under word dictionary model (WDM), The probability of generating sentence S is:

$$P(S|D, \boldsymbol{\theta}) = \sum_{k=1}^{K} \theta_{i_k}.$$

The probability of generating text T is:

$$P(T|D, \boldsymbol{\theta}) = \sum_{S \in C_T} P(S|D, \boldsymbol{\theta}).$$

Let  $T = \{T_j\}_{j=1}^n$  be the observed variable,  $S = \{S_j\}_{j=1}^n$  be the hidden random variable with  $S_j$  represents a segmenting way of  $T_j$ ,  $\theta$  be the parameters of model. The log likelihood of corpus T is:

$$L = \log P(T, S)$$

$$= \log \prod_{j=1}^{n} P(T_j, S_j | D, \boldsymbol{\theta})$$

$$= \log \prod_{j=1}^{n} P(S_j | D, \boldsymbol{\theta})$$

$$= \log \prod_{j=1}^{n} \prod_{S \in C_{T_j}} P(S | D, \boldsymbol{\theta})^{I(S_j = S)}$$

$$= \sum_{j=1}^{n} \sum_{S \in C_{T_j}} I(S_j = S) \log P(S | D, \boldsymbol{\theta}).$$

For E-step, Q function is defined as:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) = E(L|D, \boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) = \sum_{j=1}^{n} \sum_{S \in C_{T_j}} P(S|T_j; D, \boldsymbol{\theta}^{(r)}) \log \prod_{i=1}^{N} \theta_i^{n_i(S)},$$

where  $n_i(S)$  denotes the number of occurrences of  $w_i$  in sentence S. Note that  $n_i(S) = 0$  if  $w_i$  doesn't occur in S. For M-step, we update  $\theta$  using:

$$\boldsymbol{\theta}^{(r+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}), \quad s.t. \sum_{i=1}^{N} \theta_i = 1.$$

By introducing  $\lambda$ , the Lagrange function is defined as:

$$f = \sum_{j=1}^{n} \sum_{S \in C_{T_i}} P(S|T_j; D, \boldsymbol{\theta}^{(r)}) \log \prod_{i=1}^{N} \theta_i^{n_i(S)} + \lambda (1 - \sum_{i=1}^{N} \theta_i).$$

We need to find the solution of the following equations:

$$\frac{\partial f}{\partial \theta_i} = \sum_{j=1}^n \sum_{S \in C_{T_j}} P(S|T_j; D, \boldsymbol{\theta}^{(r)}) \cdot n_i(S) \cdot \frac{1}{\theta_i} - \lambda = 0$$

$$\frac{\partial f}{\partial \lambda} = 1 - \sum_{i=1}^N \theta_i = 0$$

The solution is as follows:

$$\lambda = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{S \in C_{T_{j}}} P(S|T_{j}; D, \boldsymbol{\theta}^{(r)}) \cdot n_{i}(S)$$

$$\hat{\theta}_{i} = \theta_{i}^{(r+1)} = \frac{\sum_{j=1}^{n} \sum_{S \in C_{T_{j}}} P(S|T_{j}; D, \boldsymbol{\theta}^{(r)}) \cdot n_{i}(S)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{S \in C_{T_{j}}} P(S|T_{j}; D, \boldsymbol{\theta}^{(r)}) \cdot n_{i}(S)}$$

Let  $n_i^{(r)}(T_j)$  represent the expected frequency of  $w_i$  in  $T_j$ , defined as:

$$n_i^{(r)}(T_j) = \sum_{S \in C_{T_i}} P(S|T_j; D, \boldsymbol{\theta}^{(r)}) \cdot n_i(S) = \frac{\sum_{S \in C_{T_j}} P(S|D, \boldsymbol{\theta}^{(r)}) n_i(S)}{P(T_j|D, \boldsymbol{\theta}^{(r)})},$$

and  $n_i^{(r)}(T) = \sum_{j=1}^n n_i^{(r)}(T_j)$  be the expected frequency of  $w_i$  in the whole corpus, we get:

$$\theta_i^{(r+1)} = \frac{n_i^{(r)}(T)}{\sum_{i=1}^N n_i^{(r)}(T)}.$$

Finally, words survived in the above EM algorithm can be ranked further with significance score  $\psi_i$ :

$$\psi_i = \sum_{j=1}^n \log \frac{P(T_j|D, \boldsymbol{\theta})}{P(T_j|D, \boldsymbol{\theta}_{w_i=0})}$$

$$= \sum_{j=1}^n \log \frac{\sum_{S \in C_{T_j}} P(S|D, \boldsymbol{\theta})}{\sum_{S \in C_{T_j}} I(w_i \notin S) P(S|D, \boldsymbol{\theta})}$$

$$= -\sum_{j=1}^n \log \frac{\sum_{S \in C_{T_j}} P(S|D, \boldsymbol{\theta}) - \sum_{S \in C_{T_j}} I(w_i \in S) P(S|D, \boldsymbol{\theta})}{\sum_{S \in C_{T_j}} P(S|D, \boldsymbol{\theta})}$$

$$= -\sum_{j=1}^n \log[1 - r_i(T_j)]$$

where  $\pmb{\theta}_{w_i=0}=(\theta_1,...,\theta_{i-1},0,\theta_{i+1},...,\theta_N)$  and  $r_i(T_j)$  is defined as:

$$r_i(T_j) = \frac{\sum_{S \in C_{T_j}} I(w_i \in S) P(S|D, \boldsymbol{\theta})}{\sum_{S \in C_{T_i}} P(S|D, \boldsymbol{\theta})} = \frac{\sum_{S \in C_{T_j}} I(w_i \in S) P(S|D, \boldsymbol{\theta})}{P(T_j|D, \boldsymbol{\theta})}$$

#### 3. Dynamic Programing

From last section, we have:

$$\begin{split} n_i^{(r)}(T_j) &= \frac{\sum_{S \in C_{T_j}} P(S|D, \pmb{\theta}^{(r)}) n_i(S)}{P(T_j|D, \pmb{\theta}^{(r)})} \quad (T_j = T; \text{ignore } D, \pmb{\theta}^{(r)}) \\ &= \frac{\sum_{t=1}^{\tau_L} \sum_{S_{|>t|} \in T_{|>t|}} \theta_{T_{|1:t|}} P(S_{|>t|}) [I(T_{|1:t|} = w_i) + n_i(S_{|>t|})]}{P(T)} \\ &= \frac{\sum_{t=1}^{\tau_L} \theta_{T_{|1:t|}} \sum_{S_{|>t|} \in T_{|>t|}} P(S_{|>t|}) [I(T_{|1:t|} = w_i) + n_i(S_{|>t|})]}{P(T)} \\ &= \frac{\sum_{t=1}^{\tau_L} \theta_{T_{|1:t|}} [I(T_{|1:t|} = w_i) \sum_{S_{|>t|} \in T_{|>t|}} P(S_{|>t|}) + \sum_{S_{|>t|} \in T_{|>t|}} P(S_{|>t|}) n_i(S_{|>t|})]}{P(T)} \\ &= \frac{\sum_{t=1}^{\tau_L} \theta_{T_{|1:t|}} [I(T_{|1:t|} = w_i) P(T_{|>t|}) + P(T_{|>t|}) \frac{\sum_{S_{|>t|} \in T_{|>t|}} P(S_{|>t|}) n_i(S_{|>t|})}{P(T)}]}{P(T)} \\ &= \frac{\sum_{t=1}^{\tau_L} \theta_{T_{|1:t|}} [P(T_{|>t|}) [I(T_{|1:t|} = w_i) + n_i(T| > t|)]}{P(T)} \\ &= \sum_{t=1}^{\tau_L} \rho_t [I(T_{|1:t|} = w_i) + n_i(T_{|>t|})] \end{split}$$

where 
$$\rho_t = \frac{\sum_{t=1}^{\tau_L} \theta_{T_{]1:t|}} P(T_{|>t|})}{P(T)}$$
 and  $P(T) = \sum_{S \in C_T} P(S) = \sum_{t=1}^{\tau_L} \theta_{T_{]1:t|}} P(T_{|>t|}).$ 

Similarly, we also have:

$$\begin{split} r_i(T_j) &= \frac{\sum_{S \in C_{T_j}} I(w_i \in S) P(S|D, \pmb{\theta})}{P(T_j|D, \pmb{\theta})} \quad (T_j = T; \text{ignore } D, \pmb{\theta}^{(r)}) \\ &= \frac{\sum_{t=1}^{\tau_L} \sum_{S_{|>t|} \in T_{|>t|}} \theta_{T_{|1:t|}} P(S_{|>t|}) [I(T_{|1:t|} = w_i) + I(w_i \in S_{|>t|}) I(T_{|1:t|} \neq w_i)]}{P(T|D, \pmb{\theta})} \\ &= \frac{\sum_{t=1}^{\tau_L} \theta_{T_{|1:t|}} [I(T_{|1:t|} = w_i) P(T_{|>t|}) + P(T_{|>t|}) \frac{\sum_{S_{|>t|} \in T_{|>t|}} P(S_{|>t|}) I(w_i \in S_{|>t|})}{P(T_{|>t|})} I(T_{|1:t|} \neq w_i)]}{P(T|D, \pmb{\theta})} \\ &= \sum_{t=1}^{\tau_L} \rho_t [I(T_{|1:t|} = w_i) + r_i(T_{|>t|}) I(T_{|1:t|} \neq w_i)] \end{split}$$

## References

Deng, K. et al. (2016). On the unsupervised analysis of domain-specific Chinese texts. Proceedings of the National Academy of Sciences, 113(22), 6154-6159.