

パラグラフベクトルの目的関数

$$L = \sum_d L_d$$

$$L_d = \frac{1}{T} \sum_{t=k}^T \log p(w_t | w_{t-k}, \dots, w_{t-1}),$$

$$p(w_t | w_{t-k}, \dots, w_{t-1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}},$$

$$y = b + Uh(w_{t-k}, \dots, w_{t-1}, d; W, D)$$

d : 文書または文

w_i : 文書または文中の i 番目の単語

W : 全ての単語の分散表現を表す行列

D : 全ての文章の分散表現を表す行列

T : 現在の文章に含まれる単語数

k : ウィンドウサイズ

ウィンドウ : ある単語の周辺を表す区間

p : softmax 関数により正規化された, 文脈から現在の単語が導かれることの尤度

h : 引数となる単語と文書または文のベクトルを結合したベクトルを返す関数

文ベクトルの重み付け平均の式

$$\mathbf{t}_{i_{part}} = \sum_{i_{sent}} \frac{w(x_{i_{part}}(i_{sent}))}{|\sum_{i'_{sent}} w(x_{i_{part}}(i'_{sent}))|} \mathbf{s}_{i_{sent}},$$

$$x_{i_{part}}(i_{sent}) = \frac{i_{sent}}{\#sent - 1} - \frac{i_{part}}{\#part - 1},$$

$$w(x) = \begin{cases} \frac{1}{2}(\cos(\pi|x|) + 1) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$\mathbf{s}_{i_{sent}}$: レビュー内の文ベクトル

$\mathbf{t}_{i_{part}}$: 重み付け平均された文ベクトル

i_{sent} : レビュー内の文ベクトルのインデックス

i_{part} : 重み付け平均された文ベクトルのインデックス

$\#sent$: レビュー内の文ベクトルの数

$\#part$: 重み付け平均された文ベクトルの数