Word Embedding

Analogies using Word Vectors:

$$e_{\text{"man"}} - e_{\text{"woman"}} \approx e_{\text{"king"}} - e_{word}$$
$$\text{argmax}_{word}\left(\text{sim}\left(e_{word}, e_{\text{"king"}} - e_{\text{"man"}} + e_{\text{"woman"}}\right)\right)$$

Cosine Similarity:

$$\text{sim}(u, v) = \frac{u^{\mathrm{T}} v}{\|u\|_2 \|v\|_2}$$
$$\text{sim}\left(e_{word}, e_{\text{"king"}} - e_{\text{"man"}} + e_{\text{"woman"}}\right) = \frac{{e_{word}}^{\mathrm{T}}\left(e_{\text{"king"}} - e_{\text{"man"}} + e_{\text{"woman"}}\right)}{\|e_{word}\|_2 \left\|e_{\text{"king"}} - e_{\text{"man"}} + e_{\text{"woman"}}\right\|_2}$$

Embedding Matrix:

$$e_i = E o_i$$

$\boldsymbol{E}$: the embedding matrix, shape = $\left(n_{feature}, n_{vocab}\right)$
$\boldsymbol{o_i}$: the one-hot vector of the *i*-th word, shape = $(n_{vocab}, 1)$
$\boldsymbol{e_i}$: the embedding vector of the *i*-th word, shape = $\left(n_{feature}, 1\right)$

Word2Vec

Skip-Grams Softmax:

$$\mathrm{P}(t|c) = \frac{e^{\theta_t^{\mathrm{T}} e_c}}{\sum_{j=1}^{n_{vocab}} e^{\theta_j^{\mathrm{T}} e_c}}$$

$\boldsymbol{c}$: the context word (input)
$\boldsymbol{t}$: the target word (output)
$\boldsymbol{e_c}$: the embedding vector of the context word, shape = $\left(n_{feature}, 1\right)$
$\boldsymbol{\theta_t}$: the parameter associated with the target word

Skip-Grams Loss Function:

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{n_{vocab}} y_i \log \hat{y}_i$$

Negative Sampling

context	word	target?
orange	juice	1
orange	king	0
orange	book	0

orange	the	0
orange	of	0
↑	↑	↑
c	***t***	***y***

Logistic Regression:

$$\mathrm{P}(y = 1 | c, t) = \mathrm{sigmoid}(\theta_t^{\mathrm{T}} e_c)$$

GloVe (Global Vectors for Word Representation)

Minimize:

$$\sum_{i=1}^{n_{vocab}} \sum_{j=1}^{n_{vocab}} \mathrm{f}\left(X_{ij}\right)\left(\theta_i^{\mathrm{T}} e_j + b_i + b_j' - \log X_{ij}\right)^2$$

$\boldsymbol{X_{ij}}$: the number of times of word *i* appearing in context of word *j*

Sentiment Classification