HW3 Word Embeddings

Jing Qian (jq2282)

1. Parameter search

(1) Table of the parameter search results

Algorithm	Win.	Dim.	N.s.	WordSim	BATS 1 (encyclopedic-semantics	BATS 2 (antonyms - binary)	BATS 3 (total)	MSR
word2vec	2	100	1	0.055	0.033	0.023	0.013	0.661
word2vec	2	100	5					
word2vec	2	100	15					
word2vec	2	300	1					
word2vec	2	300	5					
word2vec	2	300	15					
word2vec	2	1000	1					
word2vec	2	1000	5					
word2vec	2	1000	15					
word2vec	5	100	1					
word2vec	5	100	5					
word2vec	5	100	15					
word2vec	5	300	1					
word2vec	5	300	5					
word2vec	5	300	15					
word2vec	5	1000	1					
word2vec	5	1000	5					
word2vec	5	1000	15					
word2vec	10	100	1					
word2vec	10	100	5					
word2vec	10	100	15					
word2vec	10	300	1					
word2vec	10	300	5					
word2vec	10	300	15					
word2vec	10	1000	1					
word2vec	10	1000	5					
word2vec	10	1000	15	0.291	0.029	0.116	0.018	0.668
SVD	2	100	-					
SVD	2	300	-					
SVD	2	1000	-					
SVD	5	100	-					
SVD	5	300	-					
SVD	5	1000	-					
SVD	10	100	-					
SVD	10	300	-					
SVDs	10	1000	-					

(2) Written analysis of the results

2. Fun with objective functions.

1) (Preliminaries)

i)

$$\sigma(-x) = \frac{1}{1+e^{x}} = \frac{e^{-x}}{1+e^{-x}} = 1 - \frac{1}{1+e^{-x}} = 1 - \sigma(x).$$

ii) Using Chain rule, we could do the derivatives:

$$\frac{d}{dx}\sigma(x) = -\frac{1}{(1+e^{-x})^2}\frac{d}{dx}e^{-x} = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)(1-\sigma(x)).$$

iii) Using Chain rule, we could do the derivatives:

$$\frac{d}{dx}\log(\sigma(x)) = \frac{1}{\sigma(x)}\frac{d}{dx}\sigma(x) = \frac{1}{\sigma(x)}\sigma(x)(1-\sigma(x)) = 1 - \sigma(x).$$

2) (A simpified global objective)

i) The inner expectation in the SGNS loss function as a sum is:

$$\mathbb{E}_{c' \sim Pn(c)}[\log \sigma(-\vec{w} \cdot \vec{c}\,')] = \sum_{c' \in V_c} P_n(c') \cdot \log \sigma(-\vec{w} \cdot \vec{c}\,') = \sum_{c' \in V_c} \frac{N_{c'}}{N} \cdot \log \sigma(-\vec{w} \cdot \vec{c}\,').$$

ii) Since $\sum_{c \in V_c} N_{w,c} = N_w$, we have:

$$
\begin{aligned}
L &= \sum_{w\in V_w}\sum_{c\in V_c} N_{w,c}(\log\sigma(\vec{w}\cdot\vec{c}) + k\cdot\mathbb{E}_{c'\sim Pn(c)}[\log\sigma(-\vec{w}\cdot\vec{c}')]) \\
&= \sum_{w\in V_w}\sum_{c\in V_c} N_{w,c}\log\sigma(\vec{w}\cdot\vec{c}) + \sum_{w\in V_w}(\sum_{c\in V_c} N_{w,c})k\cdot\mathbb{E}_{c'\sim Pn(c)}[\log\sigma(-\vec{w}\cdot\vec{c}')] \\
&= \sum_{w\in V_w}\sum_{c\in V_c} N_{w,c}\log\sigma(\vec{w}\cdot\vec{c}) + \sum_{w\in V_w} N_w k\cdot\mathbb{E}_{c'\sim Pn(c)}[\log\sigma(-\vec{w}\cdot\vec{c}')] \\
&= \sum_{w\in V_w}\sum_{c\in V_c} N_{w,c}\log\sigma(\vec{w}\cdot\vec{c}) + \sum_{w\in V_w} N_w k\cdot\sum_{c'\in V_c}\frac{N_{c'}}{N}\cdot\log\sigma(-\vec{w}\cdot\vec{c}') \\
&= \sum_{w\in V_w}\sum_{c\in V_c} N_{w,c}\log\sigma(\vec{w}\cdot\vec{c}) + \sum_{w\in V_w} N_w k\cdot\sum_{c\in V_c}\frac{N_c}{N}\cdot\log\sigma(-\vec{w}\cdot\vec{c}) \\
&= \sum_{w\in V_w}\sum_{c\in V_c} N_{w,c}\log\sigma(\vec{w}\cdot\vec{c}) + \sum_{w\in V_w}\sum_{c\in V_c} k\cdot N_w\frac{N_c}{N}\cdot\log\sigma(-\vec{w}\cdot\vec{c}) \\
&= \sum_{w\in V_w}\sum_{c\in V_c}[N_{w,c}\log\sigma(\vec{w}\cdot\vec{c}) + k\cdot N_w\frac{N_c}{N}\cdot\log\sigma(-\vec{w}\cdot\vec{c})].
\end{aligned}
$$

3) (Optimizing at the local level)

i) If we take $x = \vec{w}\cdot\vec{c}$, we have:

$$
l = N_{w,c}\log\sigma(x) + k\cdot N_w\frac{N_c}{N}\cdot\log\sigma(-x).
$$

And using the derivatives from part 1), the derivative of l with respect to x is:

$$
\begin{aligned}
\frac{d}{dx}l &= N_{w,c}\frac{d}{dx}\log\sigma(x) + k\cdot N_w\frac{N_c}{N}\cdot\frac{d}{dx}\log\sigma(-x) \\
&= N_{w,c}(1-\sigma(x)) + k\cdot N_w\frac{N_c}{N}\cdot\frac{d}{dx}\log(1-\sigma(x)) \\
&= N_{w,c}(1-\sigma(x)) - k\cdot N_w\frac{N_c}{N}\cdot\sigma(x) \\
&= N_{w,c} - (k\cdot N_w\frac{N_c}{N} + N_{w,c})\cdot\sigma(x)
\end{aligned}
$$

ii) Setting $\frac{d}{dx}l = 0$, we have:

$$
\begin{aligned}
&N_{w,c} - (k\cdot N_w\frac{N_c}{N} + N_{w,c})\cdot\sigma(x) = 0, \\
&\sigma(x) = \frac{N_{w,c}}{k\cdot N_w\frac{N_c}{N} + N_{w,c}} = \frac{1}{1 + k\cdot\frac{N_w N_c}{N N_{w,c}}}, \\
&\frac{1}{1+e^{-x}} = \frac{1}{1 + k\cdot\frac{N_w N_c}{N N_{w,c}}}, \\
&x = -\log(k\cdot\frac{N_w N_c}{N N_{w,c}}) = -\log(\frac{N_w N_c}{N N_{w,c}}) - \log k = \log(\frac{N_{w,c}N}{N_w N_c}) - \log k.
\end{aligned}
$$

iii) Since $x = \vec{w} \cdot \vec{c}$, the optimal $\vec{w} \cdot \vec{c}$ corresponds to the optimal x in part 3) ii). Also, according to the definition of $PMI(w, c)$, we have:

$$\vec{w} \cdot \vec{c} = x = \log(\frac{N_{w,c}N}{N_w N_c}) - \log k = PMI(w, c) - \log k.$$