Cs224n - Assignment 2 - version 2021

Artiom Matvei

February 2024

Note

See a2.pdf file for problem descriptions.

Understanding word2vec

Question 1.a

To show that

$$\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o),$$

note that $y_w = 0$ for all w except when $w = o$ in which case $y_w = 1$. This is from the problem statement of question 1 in the assignment 2 pdf file (see 6th paragraph).

Question 1.b

Recall that

$$\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log P(O = o|C = c),$$

where

$$P(o|c) = \frac{\exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right)}$$

To compute the partial derivative of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to $\boldsymbol{v}_c$ let's first compute

$$\frac{\partial}{\partial \boldsymbol{v}_c} \log P(o|c)$$

where

$$P(o|c) = \frac{\exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right)}.$$

We get

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{v}_c}\log P(o|c) &= \frac{\partial}{\partial \boldsymbol{v}_c}\log\frac{\exp(\boldsymbol{u}_o^T\boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}}\exp\left(\boldsymbol{u}_w^T\boldsymbol{v}_c\right)}\\ &= \frac{\partial}{\partial \boldsymbol{v}_c}\log\exp(\boldsymbol{u}_o^T\boldsymbol{v}_c) - \frac{\partial}{\partial \boldsymbol{v}_c}\log\sum_{w\in Vocab}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c).\end{aligned}$$

Let's compute the derivatives of the two terms separately. The first term gets simplified because log and exp are inverse operations and because

$$\partial_x u^T x = \partial_x x^T u = u.$$

Thus we get

$$\frac{\partial}{\partial \boldsymbol{v}_c}\log\exp(\boldsymbol{u}_o^T\boldsymbol{v}_c) = \frac{\partial}{\partial \boldsymbol{v}_c}\boldsymbol{u}_o^T\boldsymbol{v}_c = \boldsymbol{u}_o.$$

Now let's compute the derivative of the second term.

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{v}_c}\log\sum_{w\in Vocab}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c) &= \frac{1}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)}\sum_{w\in\text{Vocab}}\frac{\partial}{\partial \boldsymbol{v}_c}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)\\ &= \frac{\sum_{x\in\text{Vocab}}\exp(\boldsymbol{u}_x^T\boldsymbol{v}_c)\boldsymbol{u}_x}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)}\\ &= \sum_{x\in\text{Vocab}}\frac{\exp(\boldsymbol{u}_x^T\boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)}\boldsymbol{u}_x\\ &= \sum_{x\in\text{Vocab}}P(x|c)\boldsymbol{u}_x.\end{aligned}$$

Putting the terms of the derivative together we get

$$\frac{\partial}{\partial \boldsymbol{v}_c}\log P(o|c) = \boldsymbol{u}_o - \sum_{x\in\text{Vocab}}P(x|c)\boldsymbol{u}_x \tag{1}$$

Having done this preliminary work we can go back to computing the partial derivative of the cost function by using the result we got in (1). Thus we have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{v}_c}\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= -\frac{\partial}{\partial \boldsymbol{v}_c}\log P(o|c)\\ &= -(\boldsymbol{u}_o - \sum_{x\in\text{Vocab}}P(x|c)\boldsymbol{u}_x)\\ &= -(\boldsymbol{U}\boldsymbol{y} - \boldsymbol{U}\hat{\boldsymbol{y}})\\ &= \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y})\end{aligned}$$

Note that the penultimate equality comes from the following derivation where $V = |\text{Vocab}|$,

$$\begin{aligned}
\sum_{x \in \text{Vocab}} P(x|c)\boldsymbol{u}_x &= P(O=1|C=c)\boldsymbol{u}_1 + ... \\
&\quad + P(O=x|C=c)\boldsymbol{u}_x + ... + P(O=V|C=c)\boldsymbol{u}_V \\
&= \begin{pmatrix} u_1^1 P(O=1|C=c) \\ \vdots \\ u_1^d P(O=1|C=c) \end{pmatrix} + \ldots + \begin{pmatrix} u_V^1 P(O=V|C=c) \\ \vdots \\ u_V^d P(O=V|C=c) \end{pmatrix} \\
&= \begin{pmatrix} u_1^1 P(O=1|C=c) + \ldots + u_V^1 P(O=V|C=c) \\ \vdots \\ u_1^d P(O=1|C=c) + \ldots + u_V^d P(O=V|C=c) \end{pmatrix} \\
&= \begin{pmatrix} u_1^1 & \ldots & u_V^1 \\ \vdots & & \vdots \\ u_1^d & \ldots & u_V^d \end{pmatrix} \begin{pmatrix} P(O=1|C=c) \\ \vdots \\ P(O=V|C=c) \end{pmatrix}
\end{aligned}$$

Question 1.c

We have

$$\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log P(O=o|C=c) = \log \frac{\exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\sum_{x \in \text{Vocab}} \exp(\boldsymbol{u}_x^T \boldsymbol{v}_c}.$$

Thus

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{u}_w} \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= -\frac{\partial}{\partial \boldsymbol{u}_w} \log \frac{\exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\sum_{x \in \text{Vocab}} \exp(\boldsymbol{u}_x^T \boldsymbol{v}_c)}) \\
&= -\frac{\partial}{\partial \boldsymbol{u}_w} \big[\log \exp(\boldsymbol{u}_o^T \boldsymbol{v}_c) - \log \sum_{x \in \text{Vocab}} \exp(\boldsymbol{u}_x^T \boldsymbol{v}_c) \big] \\
&= \frac{\partial}{\partial \boldsymbol{u}_w} \log \sum_{x \in \text{Vocab}} \exp(\boldsymbol{u}_x^T \boldsymbol{v}_c) - \frac{\partial}{\partial \boldsymbol{u}_w} \boldsymbol{u}_o^T \boldsymbol{v}_c
\end{aligned}$$

Now we'll compute the two terms separately by looking at the cases where $w = o$ and where $w \neq o$.

If we assume that $w \neq o$ then the 2nd term is 0 because the dot product is then constant with respect to $\boldsymbol{u}_w$. On the other hand, if we assume that $w = o$ then the second term yields $\boldsymbol{v}_c$

Now consider the 1st term,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{u}_w}\log\sum_{x\in\text{Vocab}}\exp(\boldsymbol{u}_x^T\boldsymbol{v}_c) =& \frac{1}{\sum_{x\in\text{Vocab}}\exp(\boldsymbol{u}_x^T\boldsymbol{v}_c)}\sum_{y\in\text{Vocab}}\frac{\partial}{\partial \boldsymbol{u}_w}\exp(\boldsymbol{u}_y^T\boldsymbol{v}_c)\\ =& \frac{1}{\sum_{x\in\text{Vocab}}\exp(\boldsymbol{u}_x^T\boldsymbol{v}_c)}\frac{\partial}{\partial \boldsymbol{u}_w}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)\\ =& \frac{\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)}{\sum_{x\in\text{Vocab}}\exp(\boldsymbol{u}_x^T\boldsymbol{v}_c)}\boldsymbol{v}_c\\ =& P(O=w|C=c)\boldsymbol{v}_c\end{aligned}$$

Therefore now putting the two terms together, if $w = o$ we have

$$\frac{\partial}{\partial \boldsymbol{u}_w}\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = (P(O=w|C=c)-1)\boldsymbol{v}_c = (\hat{y}_w - 1)\boldsymbol{v}_c,$$

and if $w \neq o$ we have

$$\frac{\partial}{\partial \boldsymbol{u}_w}\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = P(O=w|C=c)\boldsymbol{v}_c = \hat{y}_w\boldsymbol{v}_c.$$

Question 1.d

Let $N = |Vocab|$. Since $\boldsymbol{U}$ is a matrix of column vectors, i.e.

$$\boldsymbol{U} = \begin{pmatrix} \vdots & & \vdots \\ \boldsymbol{u}_1 & \dots & \boldsymbol{u}_N \\ \vdots & & \vdots \end{pmatrix},$$

the derivatives of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ need to be arranged in a manner to respect the shape convention. That is

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{U}}\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) =\\ &= \begin{pmatrix} \vdots & & \vdots \\ \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c,o,\boldsymbol{U})}{\boldsymbol{u}_1} & \dots & \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c,o,\boldsymbol{U})}{\boldsymbol{u}_N} \\ \vdots & & \vdots \end{pmatrix}\\ &= \begin{pmatrix} \vdots & & \vdots & & \vdots \\ \hat{y}_1\boldsymbol{v}_c & \dots & (\hat{y}_o-1)\boldsymbol{v}_c & \dots & \hat{y}_N\boldsymbol{v}_c \\ \vdots & & \vdots & & \vdots \end{pmatrix}\\ &= \begin{pmatrix} \vdots & & \vdots & & \vdots \\ (\hat{y}_1-y_1)\boldsymbol{v}_c & \dots & (\hat{y}_o-y_o)\boldsymbol{v}_c & \dots & (\hat{y}_N-y_N)\boldsymbol{v}_c \\ \vdots & & \vdots & & \vdots \end{pmatrix}\\ &= (\hat{\boldsymbol{y}}-\boldsymbol{y})^T\boldsymbol{v}_c\end{aligned}$$

Question 1.e

Let

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

First note that

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

Then

$$\begin{aligned}
\sigma'(x) &= \left(\frac{e^x}{e^x + 1}\right)' \\
&= \frac{e^x(e^x + 1) - e^x e^x}{(e^x + 1)^2} \\
&= \frac{e^x(e^x + 1 - e^x)}{(e^x + 1)^2} \\
&= \frac{e^x}{(e^x + 1)^2} \\
&= \frac{\sigma(x)}{e^x + 1} \\
&= \sigma(x)(1 - \sigma(x))
\end{aligned}$$

Question 1.f

We have

$$\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) - \sum_{k=1}^{K} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))$$

Derivative with respect to central vector

$$\frac{\partial}{\partial \boldsymbol{v}_c} \boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\frac{\partial}{\partial \boldsymbol{v}_c} \log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) - \sum_{k=1}^{K} \frac{\partial}{\partial \boldsymbol{v}_c} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))$$

Let's compute the first term,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{v}_c} \log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) &= \frac{1}{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)} \frac{\partial}{\partial \boldsymbol{v}_c} \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c) \\
&= \frac{1}{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)} \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)(1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) \boldsymbol{u}_o \\
&= (1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) \boldsymbol{u}_o.
\end{aligned}$$

Now let's compute the second term, which is very similar to the first one, thus by symmetry we get,

$$\frac{\partial}{\partial \boldsymbol{v}_c} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) = (1 - \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))(-\boldsymbol{u}_k)$$

Putting both terms together we get

$$\frac{\partial}{\partial \boldsymbol{v}_c}\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = (\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c) - 1)\boldsymbol{u}_o + \sum_{k=1}^{K}(1 - \sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c))\boldsymbol{u}_k$$

Derivative with respect to $\boldsymbol{u}_o$

$$\frac{\partial}{\partial \boldsymbol{u}_o}\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\frac{\partial}{\partial \boldsymbol{u}_o}\log(\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)) - \sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{u}_o}\log(\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c))$$

Note that the second term above is zero because every summand inside the summation is constant with respect to $\boldsymbol{u}_o$ because $o \notin \{w_1, ..., w_K\}$ as written in the problem statement. Therefore we get,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{u}_o}\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= -\frac{\partial}{\partial \boldsymbol{u}_o}\log(\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)) \\
&= -\frac{1}{\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)}\frac{\partial}{\partial \boldsymbol{u}_o}\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c) \\
&= -\frac{1}{\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)}\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)(1 - \sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c))\frac{\partial}{\partial \boldsymbol{u}_o}(\boldsymbol{u}_o^T\boldsymbol{v}_c) \\
&= (\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c) - 1)\boldsymbol{v}_c
\end{aligned}$$

Derivative with respect to a negative sample $\boldsymbol{u}_k$

In the problem statement we assumed that the K negative samples are distinct. Let $\hat{k} \in \{1, ..., K\}$ be the variable with respect to which we'll compute the variation. Therefore we have,

$$\frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}}\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}}\log(\sigma(\boldsymbol{u}_o^T\boldsymbol{v}_c)) - \sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}}\log(\sigma(-\boldsymbol{u}_k^T\boldsymbol{v}_c))$$

Note that the first term is equal to zero because the first term is constant with respect to $\boldsymbol{u}_{\hat{k}}$ as $\hat{k} \neq o$. Also note that all the terms in the second term are also equal to zero because they are also constant with respect to $\boldsymbol{u}_{\hat{k}}$ except for the only one when $k = \hat{k}$. Therefore,

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}}\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) \\
&\quad = -\frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}}\log(\sigma(-\boldsymbol{u}_{\hat{k}}^T\boldsymbol{v}_c)) \\
&\quad = -\frac{1}{\sigma(-\boldsymbol{u}_{\hat{k}}^T\boldsymbol{v}_c)}\sigma(-\boldsymbol{u}_{\hat{k}}^T\boldsymbol{v}_c)(1 - \sigma(-\boldsymbol{u}_{\hat{k}}^T\boldsymbol{v}_c))\frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}}(-\boldsymbol{u}_{\hat{k}}^T\boldsymbol{v}_c) \\
&\quad = (1 - \sigma(-\boldsymbol{u}_{\hat{k}}^T\boldsymbol{v}_c))\boldsymbol{v}_c
\end{aligned}$$

Note that negative sampling is much more efficient to compute that the naive-softmax loss because the naive-softmax contains a summation spanning over the whole vocabulary while negative sampling contains a summation over K elements only.

Question 1.g

We have

$$\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) - \sum_{k=1}^{K} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))$$

where $w_i = w_j$ may be true when $i \neq j$. Let $\hat{k} \in \{1, ..., K\}$ be the variable with respect to which we'll compute the variation. Note then that the derivative of the first term with respect to $\boldsymbol{u}_{\hat{k}}$ is equal to zero. Therefore, letting

$$A = \left| \left\{ k \in \{1, ..., K\} | w_k = w_{\hat{k}} \right\} \right|,$$

we get

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}} \boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) \\
&\quad = -\frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}} \left[\sum_{\{k | w_k = w_{\hat{k}}\}} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) + \sum_{\{k | w_k \neq w_{\hat{k}}\}} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \right] \\
&\quad = -\sum_{\{k | w_k = w_{\hat{k}}\}} \frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) - \sum_{\{k | w_k \neq w_{\hat{k}}\}} \frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \\
&\quad = -\sum_{\{k | w_k = w_{\hat{k}}\}} \frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) - \sum_{\{k | w_k \neq w_{\hat{k}}\}} 0 \\
&\quad = -A \frac{\partial}{\partial \boldsymbol{u}_{\hat{k}}} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \\
&\quad = A(1 - \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \boldsymbol{v}_c
\end{aligned}$$

Question 1.h

Skip-gram derivative with respect to $\boldsymbol{U}$

$$\begin{aligned}
\partial_{\boldsymbol{U}} \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, ...w_{t+m}, \boldsymbol{U}) &= \partial_{\boldsymbol{U}} \sum_{-m \leq j \leq m, j \neq 0} \boldsymbol{J}(\boldsymbol{v}_c, w_t + j, \boldsymbol{U}) \\
&= \sum_{-m \leq j \leq m, j \neq 0} \partial_{\boldsymbol{U}} \boldsymbol{J}(\boldsymbol{v}_c, w_t + j, \boldsymbol{U})
\end{aligned}$$

Skip-gram derivative with respect to $\boldsymbol{v}_c$

$$\begin{aligned}\partial_{\boldsymbol{v}_c} \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, ...w_{t+m}, \boldsymbol{U}) &= \partial_{\boldsymbol{v}_c} \sum_{-m \leq j \leq m, j \neq 0} \boldsymbol{J}(\boldsymbol{v}_c, w_t + j, \boldsymbol{U}) \\ &= \sum_{-m \leq j \leq m, j \neq 0} \partial_{\boldsymbol{v}_c} \boldsymbol{J}(\boldsymbol{v}_c, w_t + j, \boldsymbol{U})\end{aligned}$$

Skip-gram derivative with respect to $\boldsymbol{v}_w$ for $w \neq c$

$$\begin{aligned}\partial_{\boldsymbol{v}_w} \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, ...w_{t+m}, \boldsymbol{U}) &= \partial_{\boldsymbol{v}_w} \sum_{-m \leq j \leq m, j \neq 0} \boldsymbol{J}(\boldsymbol{v}_c, w_t + j, \boldsymbol{U}) \\ &= \sum_{-m \leq j \leq m, j \neq 0} \partial_{\boldsymbol{v}_w} \boldsymbol{J}(\boldsymbol{v}_c, w_t + j, \boldsymbol{U}) \quad = 0\end{aligned}$$

where the last line is true because $\boldsymbol{J}$ is constant with respct to $\boldsymbol{v}_w$.

Implementing word2vec

After implementing word2vec as per the instruction of the assignment 2 pdf document, I got the word embedding seen in Figure 1. As expected, antonyms group together as they could be used fairly interchangeably in the same sentence (e.g. great vs boring), and synonyms are also close together for the same reason (e.g. amazing and wonderful).

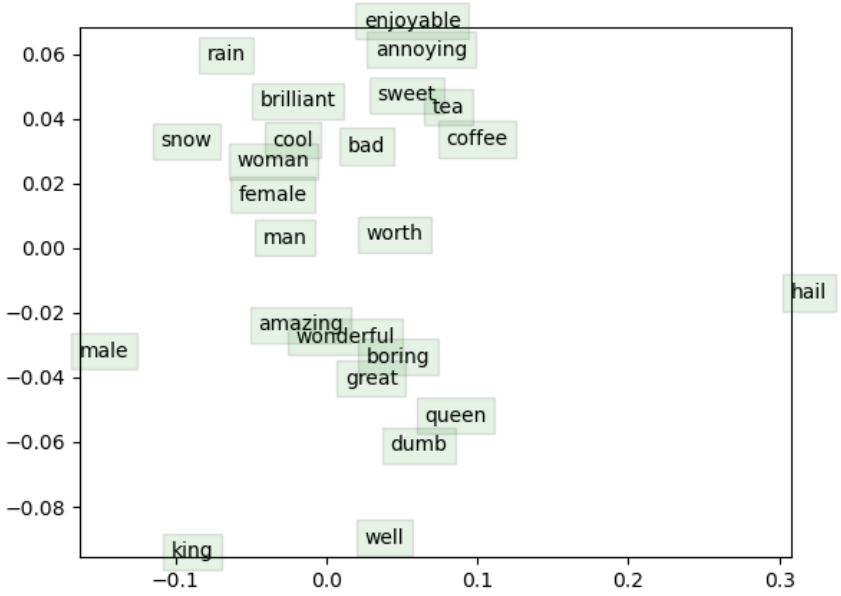

Figure 1: Word embedding after training the word2vec