Théo VIEL
theo.viel.enpc@gmail.com

Lab session # 2
ALTEGRAD 2019

10/27/19

1 Question 1

Using the sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$, the computations are straight forward and we get :

$$\frac{\partial L}{\partial w_c^+} = -w_t \sigma(-w_c^+ w_t)$$

$$\frac{\partial L}{\partial w_c^-} = w_t \sigma(w_c^- w_t)$$

2 Question 2

With the same notations :

$$\frac{\partial L}{\partial w_t} = \sum_{c \in C_t^+} -w_c \sigma(-w_c w_t) + \sum_{c \in C_t^-} w_c \sigma(w_c w_t)$$

3 Question 3

The cosine similarity between "movie" and "film" which are similar words is 0.9967 whereas between "movie" and banana it is -0.1173. Words which have the same meaning are indeed close in our representation space, and words with no relations whatsoever are far away, which is what we want.
The same thing appears in the t-SNE embedding plot. For instance, here are some words with very close representations in the 2D space :

- black, white
- better, worse
- entire, whole

Which leads us to conclude that the learnt embedding space translates pretty well the similarity in vectors distance, which is what its aim is.

4 Question 4

The idea is not only to use the word vectors in the window for prediction, but also a context vector that represents the whole document. This vector will be optimized the same way as the word vectors, to predict a hidden word in a sequence. The main difference is that every word in a document will share the same context vector.