Yassine Filali
yassine.filali@ensta-paris.fr

Lab session # LAB 1
ALTEGRAD 2020

11/24/20

1 Question 1

For one positive example, the partial derivative of the loss is :

$$\frac{\partial L_{c^+}}{\partial w_c^+} = \frac{\partial \log\left(1 + e^{-w_{c^+} \cdot w_t}\right)}{\partial w_{c^+}}$$
$$= \frac{-w_t}{1 + e^{-w_{c^+} \cdot w_t}}$$

For one negative example :

$$\frac{\partial L_{c^-}}{\partial w_c^-} = \frac{\partial \log\left(1 + e^{-w_{c^-} \cdot w_t}\right)}{\partial w_{c^-}}$$
$$= \frac{w_t}{1 + e^{-w_{c^-} \cdot w_t}}$$

2 Question 2

The partial derivative of the losst w.r.t. the target word is:

$$\frac{\partial \mathcal{L}}{\partial w_t} = \sum_{c \in \mathcal{C}_t^+} \frac{-w_c}{1 + e^{w_c \cdot w_t}} + \sum_{c \in \mathcal{C}_t^-} \frac{w_c}{1 + e^{-w_c \cdot w_t}}$$

3 Question 3

For the values of cosine similarity applied on the given examples, we obtain a value of 0.9951 between movie and film, which is reasonable since these two words are synonyms. We also obtain a value of 0.0392 between movie and banana, as these two words are not specially related.
For the plot, we can see some words that are highly related are plotted close to each other, for example camera and scene, or might and would which are basically synonyms.

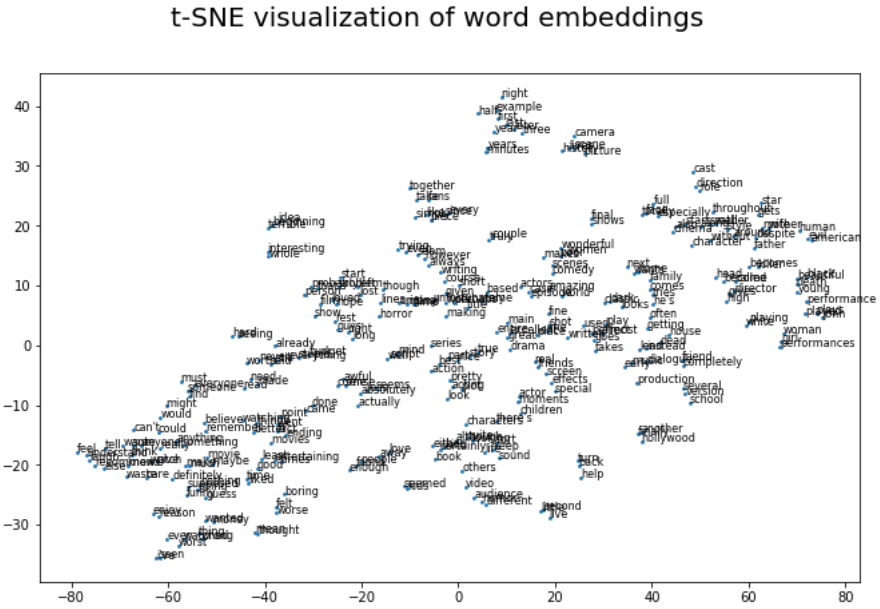

Figure 1: t-SNE visualization of word embeddings

The embedding space is of dimension 30, so there are a lot of words in the plot that are close together even though they don't appear to be related, because we only consider the two most relevant dimension.

4 Question 4

In order to learn document vectors jointly with word vector, we can add the paragraph vectors to the learning process (And the whole document is represented by a matrix, where each column represents a paragraph vector). We can use the document matrix W_d , which columns are paragraph vectors, as an input along with W_t for the learning process along the word vectors. After training, we get a matrix representation of each paragraph with the embedding matrix, and this approach takes notice of the whole context of the paragraph along the sampled windows. Another approach could be the use of just the document matrix instead of W_t to learn the context, then we can use the output matrix W_c as an embedding matrix (like the distributed bag of words in [1]). When it comes to preprocessing, we need so keep track of each document and its paragraphs.

References

[1] Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents,” 2014.