NAOUMI Salmane - KILANI Al Houceine
salmane.naoumi@gmail.com
houcinekila@gmail.com

Lab session # 2
ALTEGRAD 2019

10/29/19

1 Computing the gradient :

Considering the loss function as follows :

$$L\left(t, \mathcal{C}_t^+, \mathcal{C}_t^-\right) = \sum_{c \in \mathcal{C}_t^+} \log\left(1 + e^{-w_c \cdot w_t}\right) + \sum_{c \in \mathcal{C}_t^-} \log\left(1 + e^{w_c \cdot w_t}\right) \tag{1}$$

The partial derivatives of the loss are :

$$\begin{aligned} \frac{\partial L}{\partial w_c^+} &= \frac{\partial L}{\partial w_c^+}\left(\log\left(1 + e^{-w_c^+ \cdot w_t}\right)\right) \\ &= -\frac{w_t}{1 + e^{w_c + w_t}} \end{aligned} \tag{2}$$

$$\begin{aligned} \frac{\partial L}{\partial w_c^-} &= \frac{\partial L}{\partial w_c^-}\left(\log\left(1 + e^{w_c^- w_t}\right)\right) \\ &= \frac{w_c^-}{1 + e^{-w_c^- w_t}} \end{aligned} \tag{3}$$

$$\frac{\partial L}{\partial w_t} = \sum_{c \in C_t^+} \frac{-w_c}{1 + e^{w_c w_t}} + \sum_{c \in C_t^-} \frac{w_c}{1 + e^{-w_c w_t}} \tag{4}$$

2 Similarity values and Embedding space :

First, we compute some similarity values between different words :

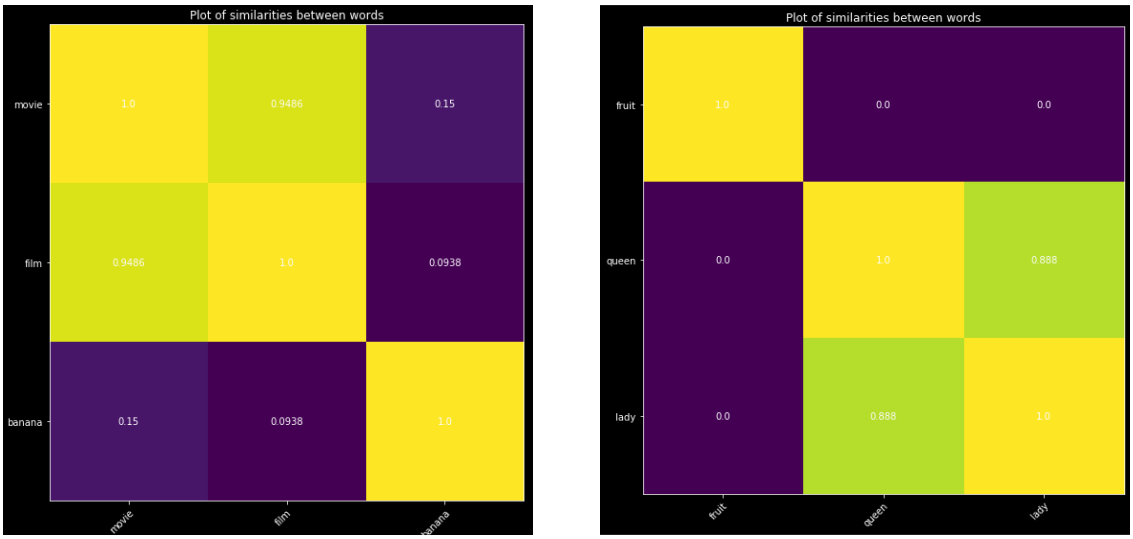

Figure 1: The generating parameter

As we can see, the skip-Gram model seems to capture relations between words such as movie and film, or queen and lady (Those words are likely to appear together). Thus, the model and loss function we used have caused some words that could occur in a same contexts to have similar space embeddings (by cosine similarity) to words.
Also, after applying PCA followed by t-Sne we get :

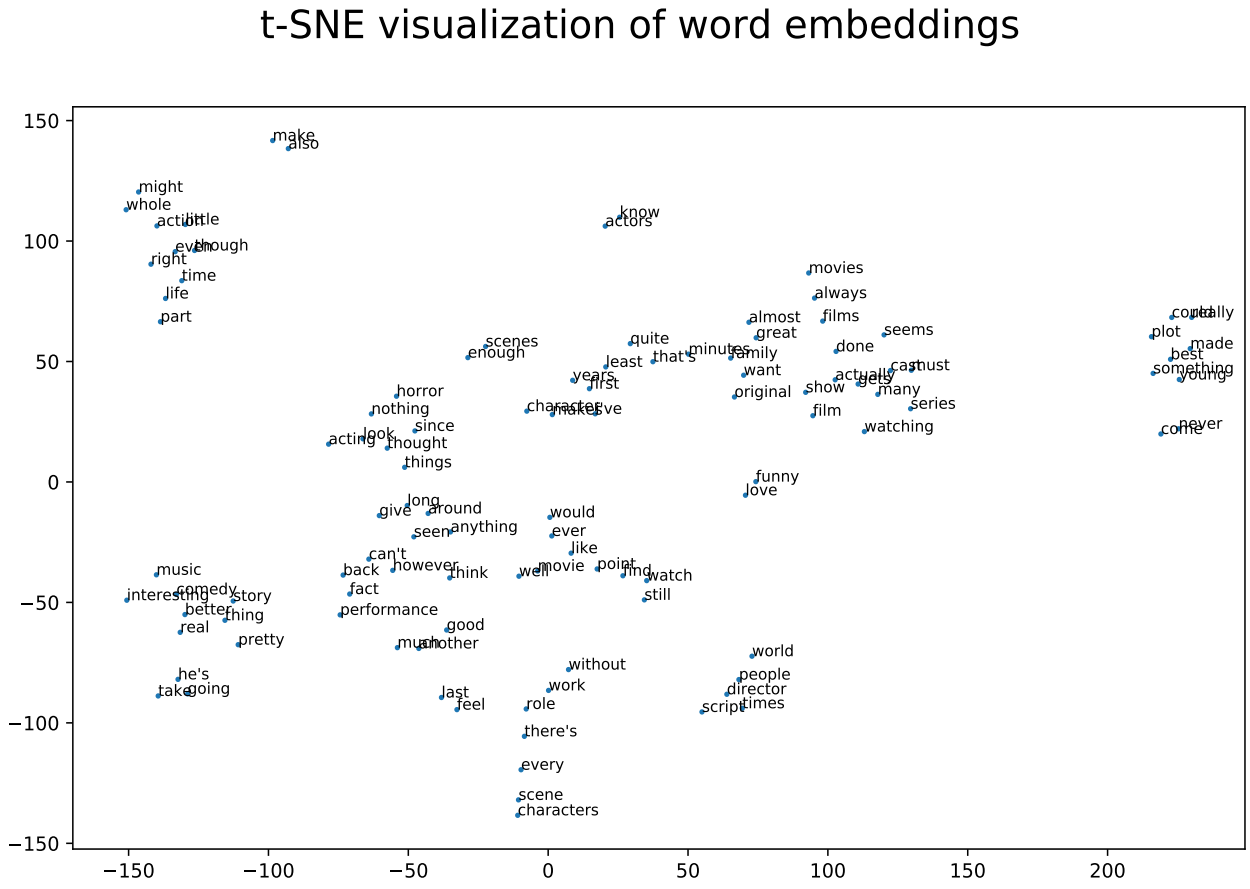

Figure 2: Learned embeddings.

We see that the words that might be used close to each others when commenting or describing a movie, are close to each other in this plot (in terms of local distances- because that's what T-SNE preserves and there is nothing to get from the "inter-clusters" distances)

3 Question 4 :

As explained in Quoc Le and Mikolov paper [1] , in order to learn document vectors jointly with word vectors, a solution is to add another input to the shallow neural network representing the id of each IMDB review. Thus, a new matrix of parameters -(named D in the paper) will contribute to the prediction of words in a context. This matrix weights will be learned by Stochastic Gradient Descent. Therefore, the training will be identical to the previous setting and the paragraph token will contribute as an identifier of the topic of a document or also act as a memory for the given context. Technically speaking, the new space embedding would be of dimension p+q (where p is the words embedding dimension and q the paragraphs embedding dimension). For the words we pad with zeros for the elements from the p+1 to the q-th position and for the paragraphs there will be zeros for the elements from the first element to the p-th element.

This representation of documents can be enriched by the use of another method that acts exactly as the previous one but the only difference is that the model will take as input the document token and will be forced to predict randomly sampled words from the document. Finally, each document will be represented as the combination of the two vectors (learned with the two different methods)

In order to compute the vector representation of a new document, we proceed by an inference step where we fix all the parameters of the model (word vectors, softmax) and apply Gradient Descent to learn the new weights of the document (document vector weights that will lead to predict the output word vector (fixed) given the context words vectors (also fixed)).

References

[1] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.