Arnaud Massenet
arnaudyoh@gmail.com

Lab session # 2
ALTEGRAD 2019

28/10/19

Question 1

$$\frac{\partial L}{\partial w_{c^+}} = \frac{-w_t \exp -w_{c^+} w_t}{1 + \exp -w_{c^+} w_t}$$

$$\frac{\partial L}{\partial w_{c^-}} = \frac{w_t \exp w_{c^-} w_t}{1 + \exp w_{c^-} w_t}$$

Question 2

$$\frac{\partial L}{\partial w_t} = \sum_{c \in C_+} \frac{-w_t \exp -w_c w_t}{1 + \exp -w_c w_t} + \sum_{c \in C_-} \frac{w_t \exp -w_c w_t}{1 + \exp -w_c w_t}$$

Question 3

It seems that the t-SNE embedding space, although not specifically designed to capture the cosine similarity, reflects well whether two words from the embedding have a high cosine similarity.

Indeed, words with high (above 0.5) cosine similarity scores tend to be closer on the plot. However, it also seems that even if two words seem far apart on our plot they could have a high cosine similarity. For example the words ”friends” at the right and ”friend” on the left of the plot in Fig. 1 are on the opposite side of the plot despite their obvious similarity.

Thus, the t-SNE embedding seems good for verifying similarities between words, but it does not seem to fare well if we want to observe dissimilarities.

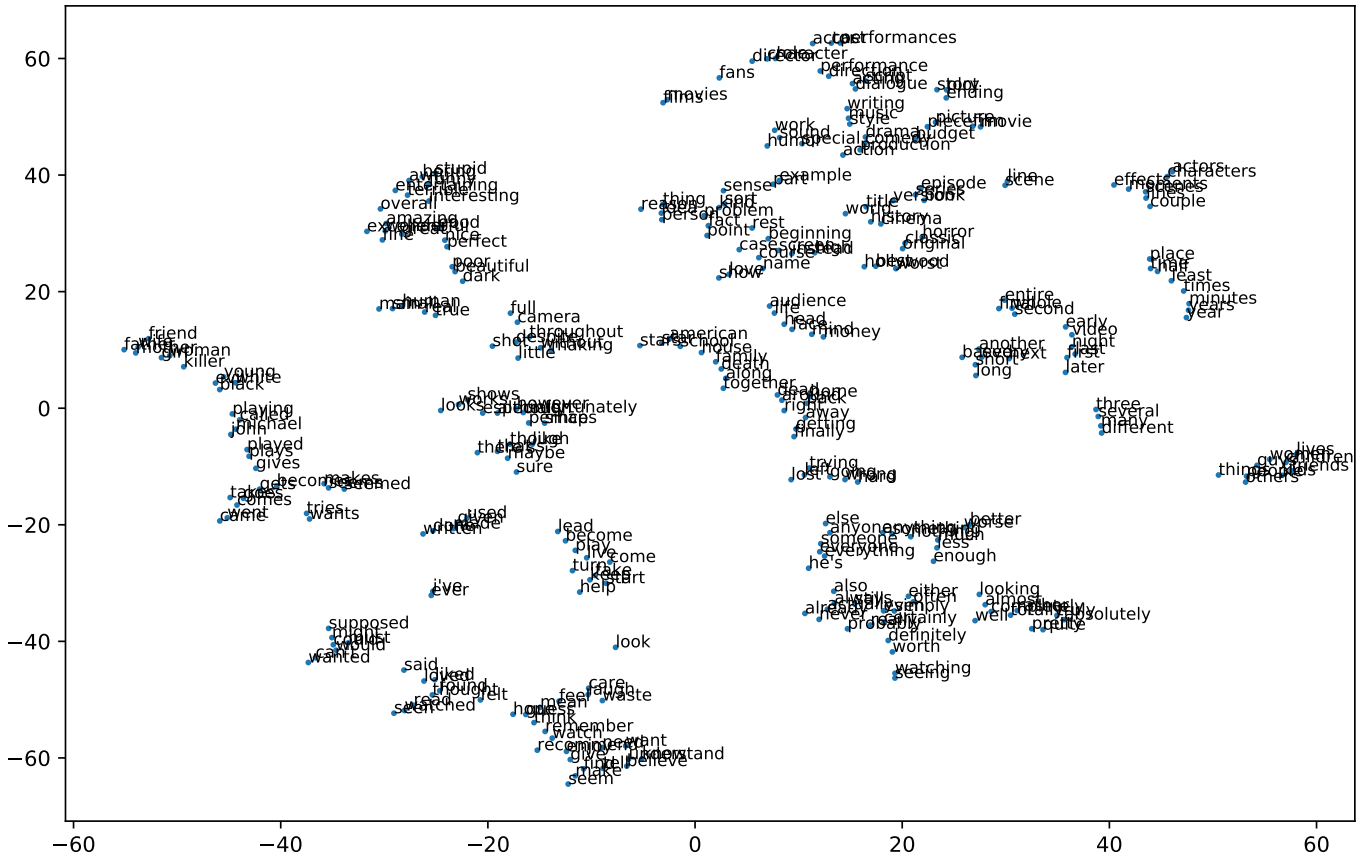

Figure 1: t-SNE word embedding of the top 500 most frequent words

Question 4

In order to capture document (paragraph) vectors along with word vectors we need to link the words and the paragraphs together.

In essence, during the preprocessing, we create for each document a list of indices to the vocabulary and we add an index that is unique to the paragraph. Thus instead of having a vocabulary of T words, we have a vocabulary of T' word/document pairs. We this new representation of the words, we could define a new objective function:

$$\arg\max_{\theta} \sum_{(t,d)\in T'} \left(\sum_{c\in C^{+}_{(t,d)}} \log p(c|(t,d);\theta) + \sum_{c\in C^{-}_{(t,d)}} \log(1-p(c|(t,d);\theta)) \right)$$

Then, during the training phase, we learn a new matrix, W_d. This matrix will hold the vector representation of each document as columns. Each document will be mapped to a unique vector. The learning of this new matrix would use Stochastic Gradient Descent using the partial derivative with respect to W_d.