HW5-Week 8-Word2Vec Embeddings based on Emotions

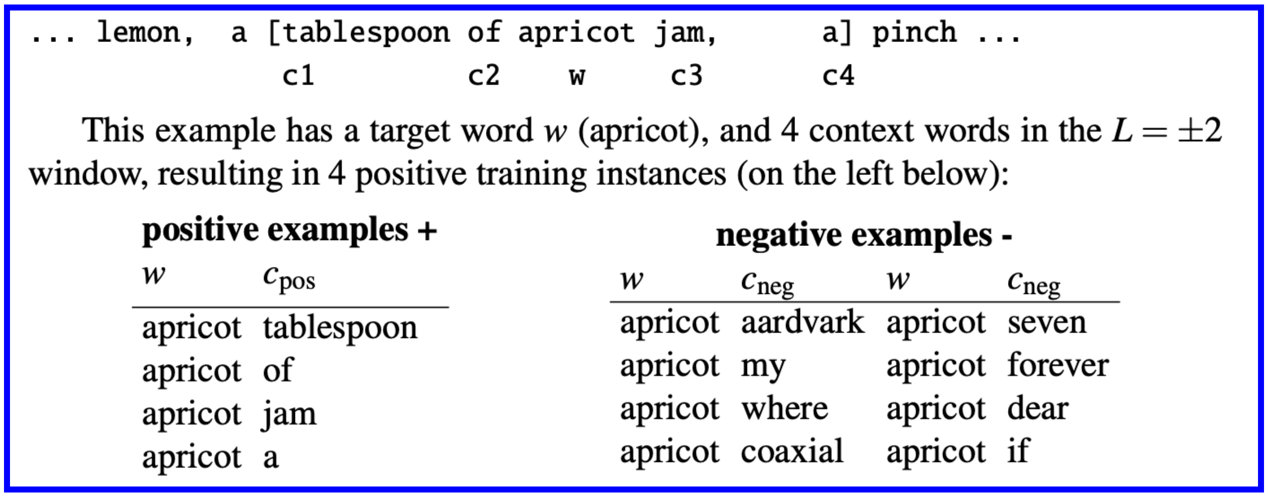
(100 points) **Discussion Week 8**
Due Sunday May 28th, 11:59 pm PST

Objective: In this assignment, you will investigate the core principles of the **Word2Vec skip-gram** algorithm.

Instructions: Complete all sections and submit the following:

1. A PDF or word doc answering the below questions.
2. Source code files **(e.g., Colab)** containing your implementation of the logistic regression classifier and data preprocessing steps.

Use the same data in the last homework with data of the 2 categories **`{Joy,Sadness}.Put all joy&sadness sentences into your corpus (there is no train/val/test split, use all rows as the training set).`** Remember the key of the implementation is that for each target word w, the neighboring words within the context window are positive words and we sample negative words. **Use context words in the L= +-2.**



```
... lemon,  a [tablespoon of apricot jam,      a] pinch ...
               c1          c2     w     c3       c4
```

This example has a target word w (apricot), and 4 context words in the $L = \pm 2$ window, resulting in 4 positive training instances (on the left below):

positive examples +

w	c_{pos}
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

negative examples -

w	c_{neg}	w	c_{neg}
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if

Section 1: Construct Training Data (10 points)

1.1: **Tokenize the text:** Start by breaking documents with multiple sentences into individual sentences, each in its own row. Then tokenize each sentence into separate words, removing punctuation in the process. **Do not stem or lemmatize the words.**

Answer:

Section in code labeled: `QUESTION 1.1` Under `SECTION 1`

1.2: **Create a dictionary:** For all the unique words, create a dictionary that maps each word to a unique integer, and vice versa.

Answer:

Section in code labeled: `QUESTION 1.2` Under `SECTION 1`

1.3: **Construct the weight and context matrices** for learning. Use **d=5**. You can use the index of word to be the index of your matrix, for example, if aardvark in your dictionary has index 1, then it is corresponding to the first row in the **target embedding W and context embedding C**. **Initialize all the weights and embeddings to be one (every element in the vector is one).**

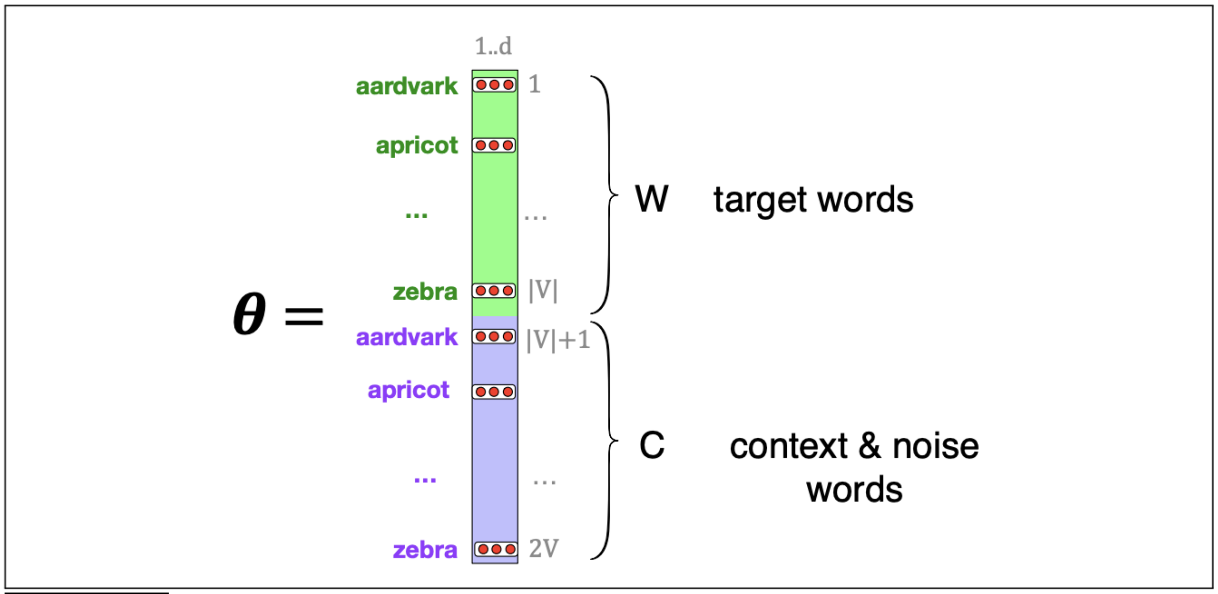


Figure 6.13 The embeddings learned by the skipgram model. The algorithm stores two embeddings for each word, the target embedding (sometimes called the input embedding) and the context embedding (sometimes called the output embedding). The parameter θ that the algorithm learns is thus a matrix of $2|V|$ vectors, each of dimension d, formed by concatenating two matrices, the target embeddings $\mathbf{W}$ and the context+noise embeddings $\mathbf{C}$.

What is your |V|?

What is the dimension of your `W` and `C` matrices?

Answer:

|V| = vocabulary_count= 1645
W Matrix = C Matrix = 1645 x 5

Section in code labeled: `QUESTION 1.3` Under `SECTION 1`

Section 2: Skip-gram with Negative Sampling (SGNS) (10 points)

For each positive word `c_pos` for the target word `w`, we’ll create `k` negative samples, each consisting of the target `w` plus a ‘noise word’ `c_neg`. Implementing this sampling function following the textbook (**Section 6.8.2**):

- The noise words are chosen according to their weighted unigram frequency `p_α(w) in your training set`, where α is a weight. Use `α = 0.75.`

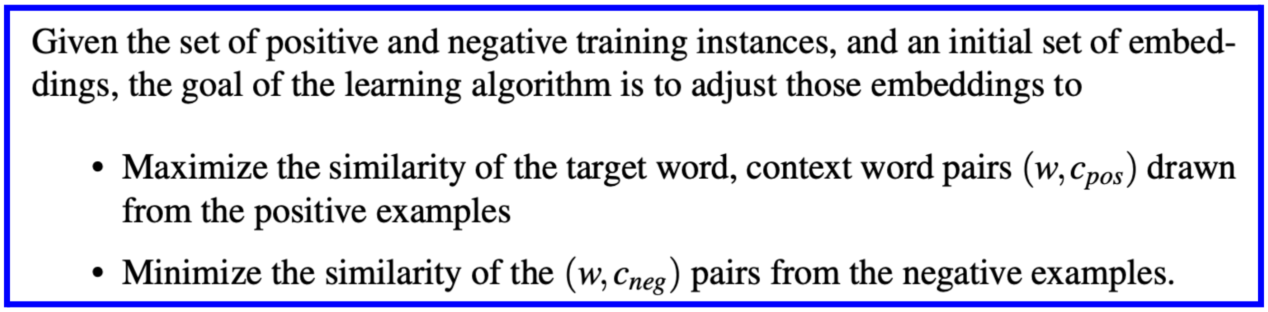
Give an example of your positive, target word pair and `k` negative pairs (k=5).

Answer:

```
1 - 'token': 'a', 'c_pos': ['the', "child's", 'abduction', 'left'],
'c_negs': ['masterpiece', 'clear', 'mike', 'immediate', 'land'].
2 - 'token': 'abandon', 'c_pos': ['she', 'decided', 'to', 'her'],
'c_negs': ['accolade', 'chemotherapy', 'i’d', 'transform', 'tell']}.
3 - 'token': 'abandoned', 'c_pos': ['the', 'child', 'looked', 'up'],
'c_negs': ['was', 'legs', 'deep', 'death', 'death']}.
```

Section in code labeled: `QUESTION 2.1` Under `SECTION 2`

Section 3: Implementing Loss Computation (30 points)


Given the set of positive and negative training instances, and an initial set of embeddings, the goal of the learning algorithm is to adjust those embeddings to

- Maximize the similarity of the target word, context word pairs (w, c_{pos}) drawn from the positive examples
- Minimize the similarity of the (w, c_{neg}) pairs from the negative examples.

Write Python code to compute the loss.

$$\frac{\partial L_{CE}}{\partial c_{pos}} = [\sigma(\mathbf{c}_{pos} \cdot \mathbf{w}) - 1]\mathbf{w} \tag{6.35}$$

$$\frac{\partial L_{CE}}{\partial c_{neg}} = [\sigma(\mathbf{c}_{neg} \cdot \mathbf{w})]\mathbf{w} \tag{6.36}$$

$$\frac{\partial L_{CE}}{\partial w} = [\sigma(\mathbf{c}_{pos} \cdot \mathbf{w}) - 1]\mathbf{c}_{pos} + \sum_{i=1}^{k}[\sigma(\mathbf{c}_{neg_i} \cdot \mathbf{w})]\mathbf{c}_{neg_i} \tag{6.37}$$

Answer:

Section in code labeled: `QUESTION 3.1` Under `SECTION 3`

Section 4: Learning & Optimization (30 points)

4.1: Complete the implementation for your SGD optimizer with different **learning rates = [0.00001, 0.0001, 0.001, 0.01, 0.1]** and fix the negative sampling k=5. What is your best learning rate and what is the lowest **training loss**? List all learning rate and loss here.

$$\mathbf{c}_{pos}^{t+1} = \mathbf{c}_{pos}^{t} - \eta[\sigma(\mathbf{c}_{pos}^{t} \cdot \mathbf{w}^{t}) - 1]\mathbf{w}^{t} \tag{6.38}$$

$$\mathbf{c}_{neg}^{t+1} = \mathbf{c}_{neg}^{t} - \eta[\sigma(\mathbf{c}_{neg}^{t} \cdot \mathbf{w}^{t})]\mathbf{w}^{t} \tag{6.39}$$

$$\mathbf{w}^{t+1} = \mathbf{w}^{t} - \eta\left[[\sigma(\mathbf{c}_{pos} \cdot \mathbf{w}^{t}) - 1]\mathbf{c}_{pos} + \sum_{i=1}^{k}[\sigma(\mathbf{c}_{neg_i} \cdot \mathbf{w}^{t})]\mathbf{c}_{neg_i}\right] \tag{6.40}$$

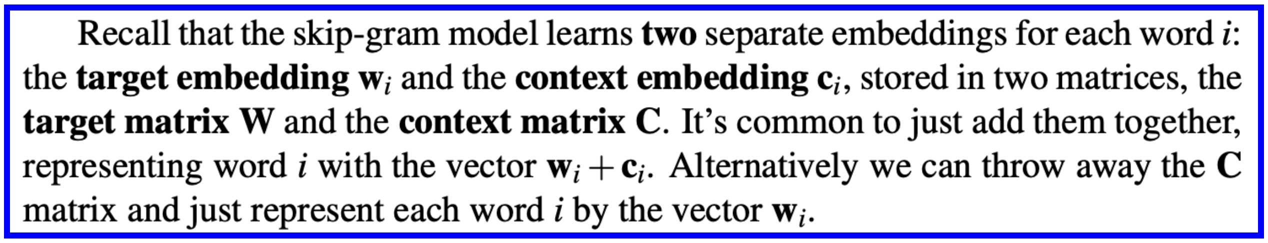
Every word in your training corpus should be utilized as a target word. Therefore, your training algorithm should iterate over every token in each sentence, using the token and its neighboring words within the context window for learning.

Answer:
```
1. Best Learning Rate: 0.1
2. Lowest Loss: 0.6371354596250903  for learning rate: 0.1
3. List of all Learning Rates and Losses:
   Learning Rate: 0.00001 Loss: 25.015838720875838
   Learning Rate: 0.00010 Loss: 24.771434262076983
   Learning Rate: 0.00100 Loss: 22.33998515256443
   Learning Rate: 0.01000 Loss: 4.644739222177062
   Learning Rate: 0.10000 Loss: 0.6395537334873087
```

Section in code labeled: `QUESTION 4.1` Under `SECTION 1`

Section 5: Final representation (10 points)

Add your W and C matrices for the final representation of all words in your vocab.



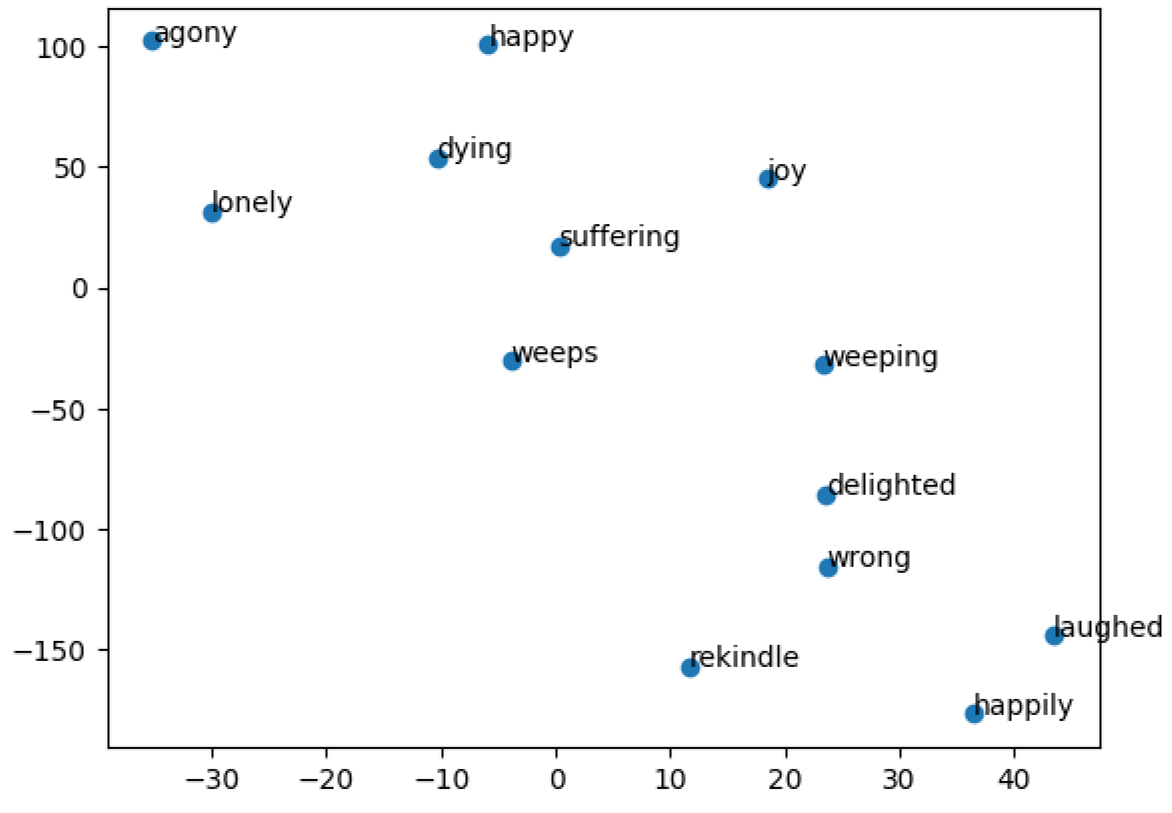
Recall that the skip-gram model learns **two** separate embeddings for each word i: the **target embedding** $\mathbf{w}_i$ and the **context embedding** $\mathbf{c}_i$, stored in two matrices, the **target matrix** $\mathbf{W}$ and the **context matrix** $\mathbf{C}$. It's common to just add them together, representing word i with the vector $\mathbf{w}_i + \mathbf{c}_i$. Alternatively we can throw away the $\mathbf{C}$ matrix and just represent each word i by the vector $\mathbf{w}_i$.

Using t-SNE (dimensionality reduction with t-distributed stochastic neighbor embedding) to reduce your word embeddings from 5 to 2. Draw the visualization for the following words: `{happy, happily, delighted, rekindle, laughed, joy, wrong, lonely, dying, agony, suffering, weeps, weeping}`

Copy and paste your plot here. You can find use existing package for the plot:

```
import matplotlib.pyplot as plt
from sklearn.manifold import TSNE
```

Answer:



Section in code labeled: `SECTION 5`