Name: Ronaldlee Ejalu
Course Name: CSC 583
Assignment number: HW#5

1. Part 1:
   - My Development environment was Google Colab and the libraries and packages I used were csv, re, os, numpy, pandas and pairwise metric
   - Results of Part 1 were:
   - ➔ Question 1
     1. word similar to dog were cat as shown below:

	similar-words	cosine_value
5448	dog-cat	0.879808

     2. words similar to whale was shark as shown below:

	similar-words	cosine_value
9860	whale-shark	0.784017

     3. words similar to before was again:

	similar-words	cosine_value
375	before-again	0.858747

     4. words similar to however was although as shown below:

	similar-words	cosine_value
375	however-although	0.965755

     5. words similar to fabricate was invent and this is shown below:

	similar-words	cosine_value
24072	fabricate-invent	0.704018

   - ➔ Question 2

```
a. dog : puppy :: cat : ?
```

```
dog : puppy :: cat : animal
```

```
b. speak : speaker :: sing : ?
```

```
speak : speaker :: sing : sang
```

```
c. France : French :: England : ?
```

```
France : French :: England : scotland
```

```
d. France : wine :: England : ?
```

```
france : wine :: england : britain
```

```
The results on part Question 2(d) surprised me because I
expected to see something like beer or a type of drink enjoyed by the
British.
```

Part 11

- My Development environment was Google Colab and the libraries and packages I used were torch, nn, optin, numpy, pandas, os, nltk, RegexpTokenizer, word_tokenize, stopwords and itertools.
- Results of Task 1 were:

```
The vocabulary size is 49.
[238.216938495636, 233.29338455200195, 228.56411004066467, 224.01450490951538, 219.6376404762268, 215.42263841629028, 211.3627860546112, 207.45027327537537, 203.67698216
The embedding vector for 'procecess' is:
tensor([ 1.6748,  0.0084, -0.7067, -0.1843, -0.9960, -0.8317, -0.4584, -0.5617,
         0.3960, -0.9836], grad_fn=&lt;SelectBackward0&gt;)
```

- The top three words that are closest to processes are our, programs and abstract as shown below:

The top three words that are closest to 'processes' by cosine similarity:

	similar-words	cosine_value
46	processes-our	0.624608
36	processes-programs	0.545113
12	processes-abstract	0.525402

- Task 2 failed to run as I was trying to run the model:

```
IndexError                                Traceback (most recent call last)
<ipython-input-25-ee02ded0be91> in <cell line: 9>()
     32         else:
     33           #print(log_probs.shape, torch.tensor([word_to_ix[target]], dtype=torch.long).shape)
---> 34           loss = loss_function(log_probs, torch.tensor([word_to_ix[target]], dtype=torch.long))
     35 
     36           # Step 5. Do the backward pass and update the gradient

2 frames
/usr/local/lib/python3.10/dist-packages/torch/nn/functional.py in nll_loss(input, target, weight, size_average, ignore_index, reduce, reduction)
   2702     if size_average is not None or reduce is not None:
   2703         reduction = _Reduction.legacy_get_string(size_average, reduce)
-> 2704     return torch._C._nn.nll_loss_nd(input, target, weight, _Reduction.get_enum(reduction), ignore_index)
   2705 
   2706 

IndexError: Target 32 is out of bounds.
```

I became a bit disappointed because I couldn't proceed past this step.

The top three words closest to titanic were brock, comes and could as shown:

The top three words that are closest to 'titanic' by cosine similarity:

	similar-words	cosine_value
646	titanic-brock	0.600329
115	titanic-comes	0.574387
162	titanic-could	0.442980

The top three words closest to acting were real, hollywood and footage as shown below:

```
The top three words that are closest to 'acting' by cosine similarity:
```

	similar-words	cosine_value
312	acting-real	0.445069
421	acting-hollywood	0.444176
69	acting-footage	0.425262

The top three words closest to great were films, begins and city as shown:

```
The top three words that are closest to 'great' by cosine similarity:
```

	similar-words	cosine_value
194	great-films	0.533587
84	great-begins	0.466959
11	great-city	0.436753

The top three words closest to poor were director, action and long as shown:

```
The top three words that are closest to 'poor' by cosine similarity:
```

	similar-words	cosine_value
379	poor-director	0.481924
452	poor-action	0.473596
269	poor-long	0.443895

General reflections about this assignment were learning how to prepare the data in the right format, the CBOW model expects it.
The difficulty I encountered was on Part 2 Task 2 with the error I encountered;
I tried to debug it but I couldn't succeed.