張雅婷

A. Total Number of Source Titles: 883,478 (排除 Re:及 Fw:開頭的標題)

Total Number of Tokenized Titles: 882,994

B. If A and B are different, what have you done for that?

排除連接詞、介詞、結構助詞(的)、標點符號、外文詞(英文)

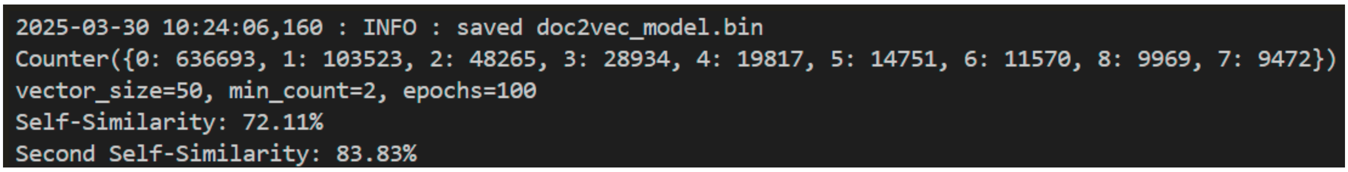
排除沒有 token 只有 label 情形，例: 英文標題

C-1. Parameters of Doc2Vec Embedding Model. (A)

a. Total Number of Training Documents: 882,994

b. Output Vector Size: 50 Min Count: 2 Epochs: 100

c. First Self Similarity: 72.11% Second Self Similarity: 83.83 %



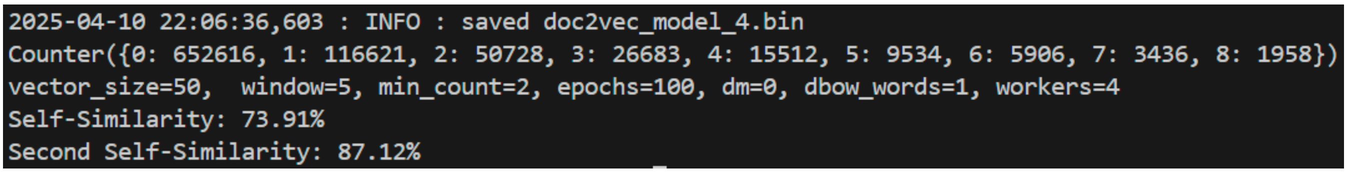
```
2025-03-30 10:24:06,160 : INFO : saved doc2vec_model.bin
Counter({0: 636693, 1: 103523, 2: 48265, 3: 28934, 4: 19817, 5: 14751, 6: 11570, 8: 9969, 7: 9472})
vector_size=50, min_count=2, epochs=100
Self-Similarity: 72.11%
Second Self-Similarity: 83.83%
```

C-2. Parameters of Doc2Vec Embedding Model. (E)

a. Total Number of Training Documents: 882,994

b. Output Vector Size: 50 Window: 5 Min Count: 2 Epochs: 100 dm: 0 (DBOW) dbow_words: 1 (訓練詞向量) Workers: 4

c. First Self Similarity: 73.91% Second Self Similarity: 87.12 %



```
2025-04-10 22:06:36,603 : INFO : saved doc2vec_model_4.bin
Counter({0: 652616, 1: 116621, 2: 50728, 3: 26683, 4: 15512, 5: 9534, 6: 5906, 7: 3436, 8: 1958})
vector_size=50,  window=5, min_count=2, epochs=100, dm=0, dbow_words=1, workers=4
Self-Similarity: 73.91%
Second Self-Similarity: 87.12%
```

D. Parameters of Multi-Class Classification Model.

a. Arrangement of Linear Layers: 50x32X16x9

b. Activation Function for Hidden Layers: ReLU

c. Activation Function for Output Layers: Softmax

d. Loss Function: Categorical Cross Entropy (nn.CrossEntropyLoss)

e. Algorithms for Back-Propagation: SGD (Stochastic Gradient Descent)

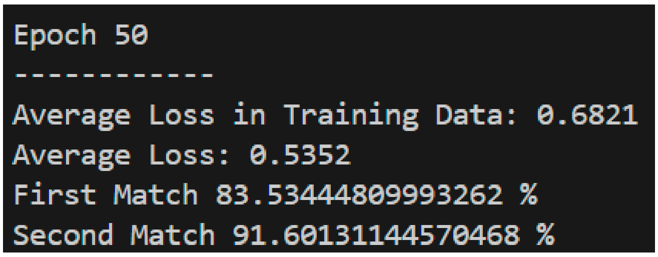
f. Total Number of Training Documents: 0.8*882,994

g. Total Number of Testing Documents: 0.2*882,994

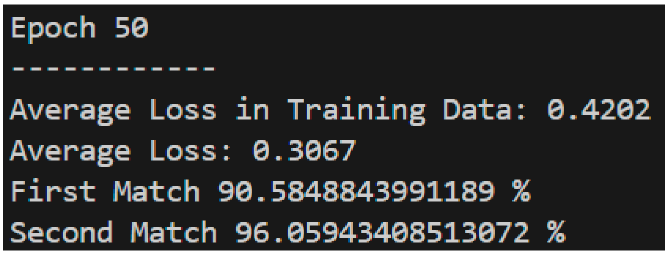
h. Epochs: 50 Learning Rate: 0.001

i-1. With Embedding (A), First Match: 83.53% Second Match: 91.60%

i-2. With Embedding (E), First Match: 90.58% Second Match: 96.06%



```
Epoch 50
------------
Average Loss in Training Data: 0.6821
Average Loss: 0.5352
First Match 83.53444809993262 %
Second Match 91.60131144570468 %
```



```
Epoch 50
------------
Average Loss in Training Data: 0.4202
Average Loss: 0.3067
First Match 90.5848843991189 %
Second Match 96.05943408513072 %
```

j. Any other parameters you think are important. 有使用 Dropout(0.2)，但差異不大

E. Parameters of Multi-Label Classification Model.

a. Arrangement of Linear Layers: 50x32x16x9

b. Activation Function for Hidden Layers: ReLU

c. Activation Function for Output Layers: Sigmoid

d. Loss Function: Binary Cross Entropy (nn.BCEWithLogitsLoss)

e. Algorithms for Back-Propagation: SGD (Stochastic Gradient Descent)

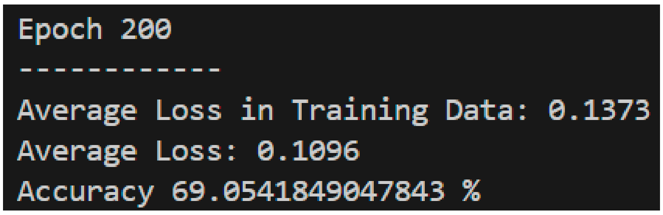
f. Total Number of Training Documents: 0.8*882,994

g. Total Number of Testing Documents: 0.2*882,994

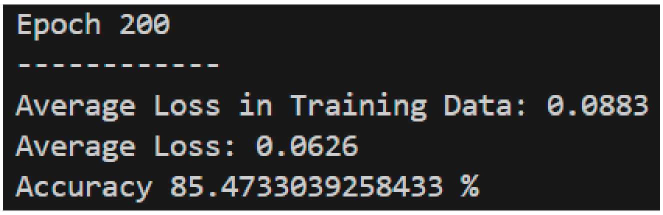
h. Epochs: 200 Learning Rate: 0.001

i-1. With Embedding (A), Threshold for Positive Label: 0.5 Accuracy Rate: 69.05%

i-2. With Embedding (E), Threshold for Positive Label: 0.5 Accuracy Rate: 85.47%



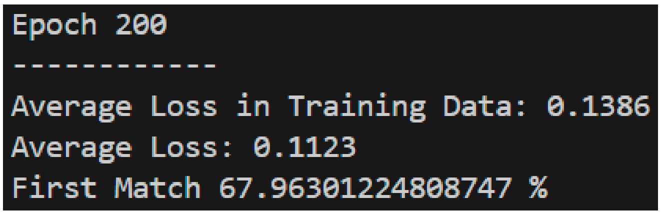
```
Epoch 200
------------
Average Loss in Training Data: 0.1373
Average Loss: 0.1096
Accuracy 69.0541849047843 %
```



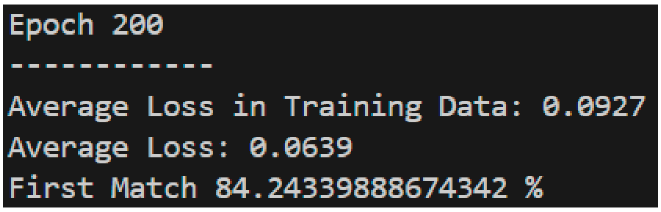
```
Epoch 200
------------
Average Loss in Training Data: 0.0883
Average Loss: 0.0626
Accuracy 85.4733039258433 %
```

i-3. With Embedding (A), Using the highest predicted probability with threshold for Positive Label: 0.5 Accuracy Rate: 67.96%

i-4. With Embedding (E), Using the highest predicted probability with threshold for Positive Label: 0.5 Accuracy Rate: 84.24%



```
Epoch 200
------------
Average Loss in Training Data: 0.1386
Average Loss: 0.1123
First Match 67.96301224808747 %
```



```
Epoch 200
------------
Average Loss in Training Data: 0.0927
Average Loss: 0.0639
First Match 84.24339888674342 %
```

j. Any other parameters you think are important. 有使用 Dropout(0.2)相較沒使用時準確率下降

F. Share your experience of optimization, including at least 2 change/result pairs.

Embedding	(A) vector_size=50 min_count=2 epochs=100	window=5 dm=1 dm_mean=1 workers=4 其餘同(A)	dm=1 workers=4 其餘同(A)	dm=0 dbow_words=1 workers=4 其餘同(A)	(E) window=5 dm=0 dbow_words=1 workers=4 其餘同(A)
Self-Similarity	72.11%	71.06%	72.75%	72.98%	73.91%
Second Self-Similarity	83.83%	83.71%	84.51%	86.66%	87.12%

Multi-class	學習率 0.001 epochs = 50	學習率 0.01 epochs = 50	Dropout(0.2) 學習率 0.001 epochs = 50	Dropout(0.2) 學習率 0.001 epochs = 50	CrossEntropyLoss (weights) 學習率 0.001 epochs = 50
Embedding Model	(A)	(A)	(A)	(E)	(A)
First Match	83.78%	85.60%	83.53%	90.58%	82.18%
Second Match	92.28%	93.41%	91.60%	96.06%	91.02%

Multi-label	學習率 0.001
	epochs = 200
	Dropout(0.2)
		學習率 0.001
		epochs = 200
		Dropout(0.2)

Embedding Model	(A)	(E)
Accuracy	69.05%	85.47%

Multi-label
Using the highest probability
	學習率 0.001
	epochs = 200
		學習率 0.01
		epochs = 200
			學習率 0.001
			epochs = 200
			Dropout(0.2)
				學習率 0.001
				epochs = 200
				Dropout(0.2)

Embedding Model	(A)	(A)	(A)	(E)
First Match	75.87%	80.71%	67.96%	84.24%