**張雅婷**

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

Total Number of Tokenized Titles: 882,994

1. 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)

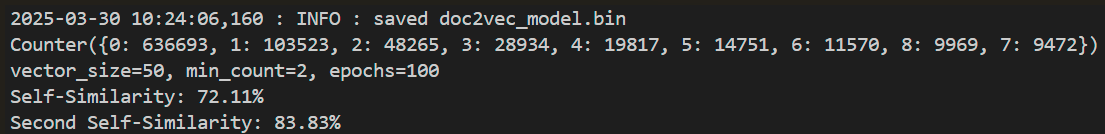
排除沒有token只有label情形，例: 英文標題

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 %

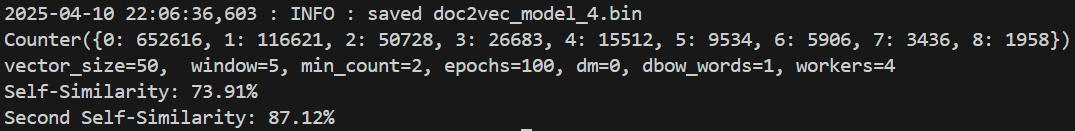


C-2. Parameters of Doc2Vec Embedding Model. (E)

a. Total Number of Training Documents: 882,994

b. Output Vector Size: 50 Window: 5 Min Count: 2 Epochs: 100 dm: 0 (DBOW忽略上下文) dbow\_words: 1 (訓練詞向量) Workers: 4

c. First 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)

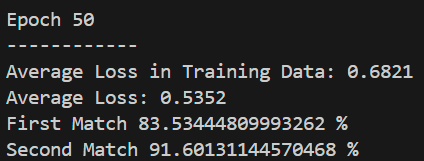
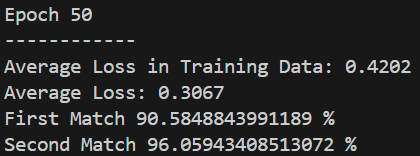
f. Total Number of Training Documents: 0.8\*882,994

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%

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)

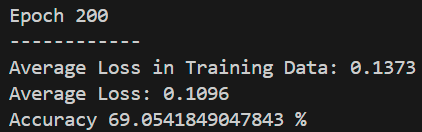
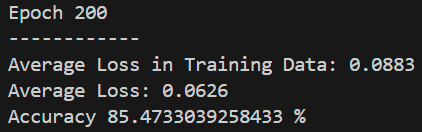
f. Total Number of Training Documents: 0.8\*882,994

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

h. Epochs: 200 Learning Rate: 0.001

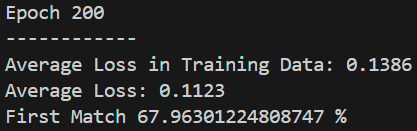
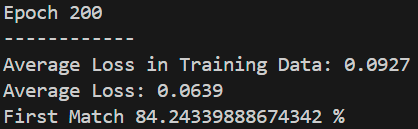
i-1. With Embedding (A), Threshold for Positive Label: 0.5 Accuracy Rate: 69.05%

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

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%

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  (預設dm:1) | window=5 dm=1(使用上下文) dm\_mean=1  (上下文詞平均向量) 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 (忽略上下文) dbow\_words=1  (訓練詞向量) 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% |

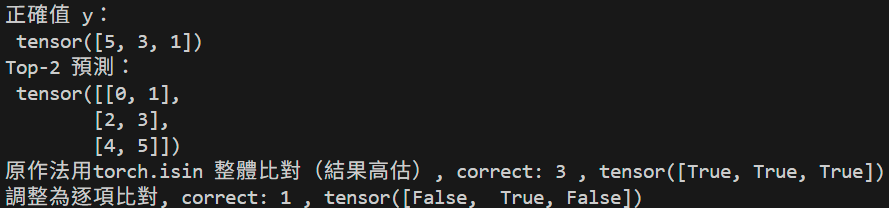
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Multi-class** | 學習率 0.001 epochs = 50 | 學習率 0.01 epochs = 50 | Dropout(0.2) 學習率 0.001 epochs = 50 | Dropout(0.2) 學習率 0.001 epochs = 50 | CrossEntropyLoss (weights) 學習率 0.001 epochs = 50 |
| Embedding Model | (A) | (A) | (A) | (E) | (A) |
| 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% |

# torch.isin（會將y比對整體 top2\_pred，而非逐項比對top2\_pred細項）

correct\_top2\_isin = torch.isin(y, top2\_pred).sum().item()





Loss 無加權

