A) Total Number of Source Titles: 1079429
Total Number of Tokenized Titles: 804998

B) If A and B are different, what have you done for that?
   a) I removed article titles starting with “re:” and “fw:” and “公告”.
   b) I removed all the punctuations, conjunctions, prepositions, and structural particles.

C) Parameters of Doc2Vec Embedding Model.
   a) Total Number of Training Documents: 804998
   b) Output vector size: 200 / Min Count: 2 / Epochs: 200 / Workers: 1
   c)First Self Similarity: 70.1% / Second Self Similarity: 72.9%

D) Parameters of Muti-Class Classification Model.
   a) Arrangement of Linear Layers: 300 × 128 × 32 × 9
   b) Activation Function for Hidden Layers: ReLU
   c)Activation Function for Output Layers: Softmax
   d) Loss Function: Categorical Cross Entropy
   e) Algorithms for Back-Propagation: Adam optimizer (adaptive moment estimation)
   f) Total Number of Training Documents: 643998
   g) Total Number of Testing Documents: 161000
   h) Epochs: 50 / Learning Rate: 0.001
   i) First Match: 83.88% Second Match: 93.26%

E) Share your experience of optimization, including at least 2 change/result pairs.
   a) Change: Removed article titles starting with "公告" from the dataset.
      Result: Accuracy remained largely unchanged, but vector inference time was reduced by about half a day.
   b) Change: Increased the number of training epochs from 30 to 50.
      Result: The test accuracy slightly improved from 83.83% to 83.88%.