A. Total Number of Source Titles: **288326**

Total Number of Tokenized Titles: **288321**

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

I removed the title with "na" (I had previously done ETL processing to handle titles with terms like "re", etc.).

C. Parameters of Doc2Vec Embedding Model(First training)

a. Total Number of Training Documents: **288321**

b. Output Vector Size: **100** , Window: **5**, Min Count: **2**, Epochs: **30** Workers: **8**

c. First Self Similarity: **79.7%** Second Self Similarity: **82.7%**

D. Parameters of Multi-Class Classification Model.

a. Arrangement of Linear Layers: **100x64x9**

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**

f. Total Number of Training Documents: **230656**

g. Total Number of Testing Documents: **57665**

h. Epochs: 30 Learning Rate: **0.001**

i. First Match: **73.02%**

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

1. Multi-Class Classification Model

I. Change epoch from 30 -> 100.

Outcome: not significant.

2. Eembedding model Change:

I. Change vector size of embedding model from 100 -> 50

Outcome: not significant .

II. Change min count of embedding model from 2 -> 5

Outcome: significant， self Similarity 79.7 -> 83

III. Change embedding model structure

a. Train two Doc2Vec models(second training).

One is PV-DM, the other is PV-DBOW. Take 50 vector size from each

model to make a mixed model.

b. Total Number of Training Documents: **288321**

c. Output Vector Size: **100** , Window: **5,** Min Count: **2**, Epochs: **30** Workers: **8**

d. First Self Similarity: **97.3%** Second Self Similarity: **99.1%**

Parameters of Multi-Class Classification Model.

a. Arrangement of Linear Layers: **100x64x9**

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**

f. Total Number of Training Documents: **230656**

g. Total Number of Testing Documents: **57665**

h. Epochs: **30** Learning Rate: **0.001**

i. First Match: **89.25%**