**Quora Question Pairs**

Abstract

Quora is a place to gain and share knowledge about anything. It’s a website to ask questions and connect with people who contribute unique insights and quality answers. Over 100 million people visit Quora every month, it's no surprise that many people ask similarly worded questions. But so many questions cause a lot of same questions with different word or different way to ask. These multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. So, we try using the NLP technique to identify duplicate questions that can provide a better experience to active seekers and writers and offer more value to both groups in the long term.

Dataset for training 404289 data

A screenshot of a question

Description automatically generated

Dataset for testing 3563490 data

A screenshot of a questionnaire

Description automatically generated

Training data:

The length of the sentence in the data

A graph with numbers and lines

Description automatically generated

The duplicate and different data comparison

A blue and orange pie chart with Crust in the background

Description automatically generated

Any common words between the questions

A graph of a number of data

Description automatically generated

Any shared words between the questions

Blue: is duplicate Red: Not duplicateA graph of a graph

Description automatically generated with medium confidence

Bert-base-uncased model:

Pretrained model on English language using a masked language modeling (MLM) objective. This model is uncased: it does not make a difference between english and English.

BERT is a transformers model pretrained on a large corpus of English data which was pretrained on the raw texts only, without humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts.

Masked language modeling (MLM): taking a sentence, the model randomly masks 15% of the words in the input then run the entire masked sentence through the model and has to predict the masked words. It allows the model to learn a bidirectional representation of the sentence. The model learns an inner representation of the English language that can then be used to extract features useful for downstream task.

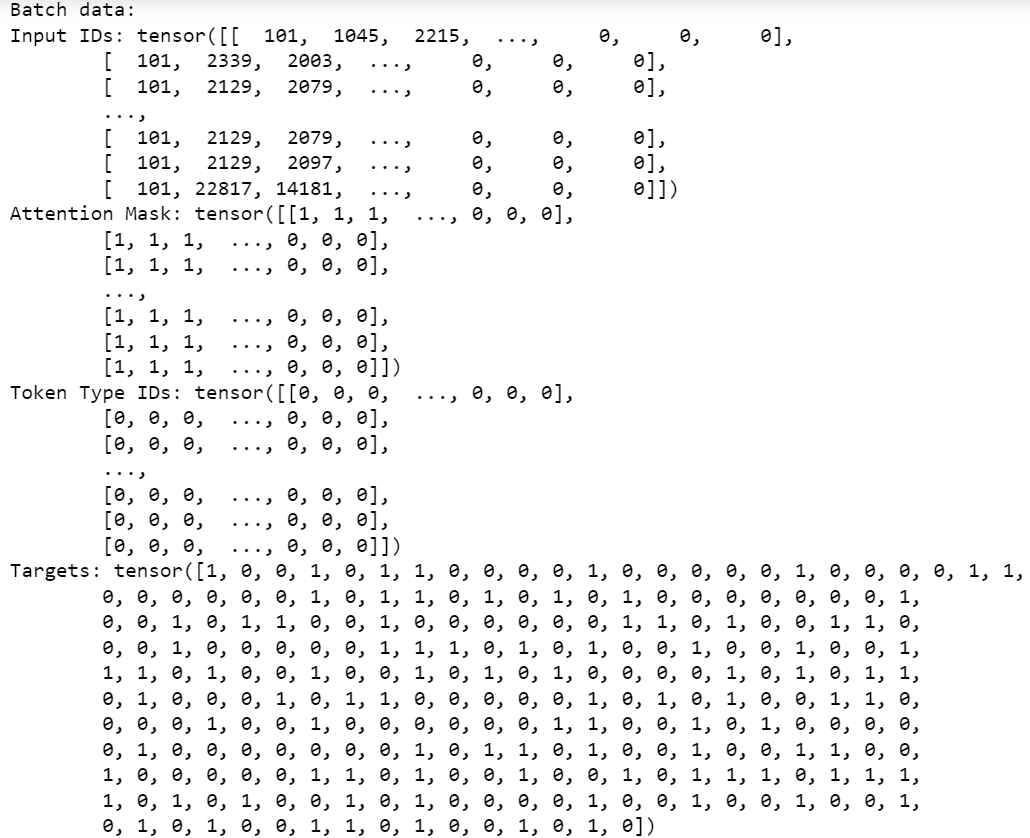
The inputs of the model

[CLS] Question1 [SEP] Question2 [SEP]

A diagram of a sentence

Description automatically generated

Transfer the data into



Training

5 epoch and Learning Rate(lr) is 3e-5 and the batch is depend on the data, it'll be 1422 in this case.

A table of numbers with black text

Description automatically generated

Prediction on test dataset

batch size as 512. And I use sigmoid to show the result of my prediction.

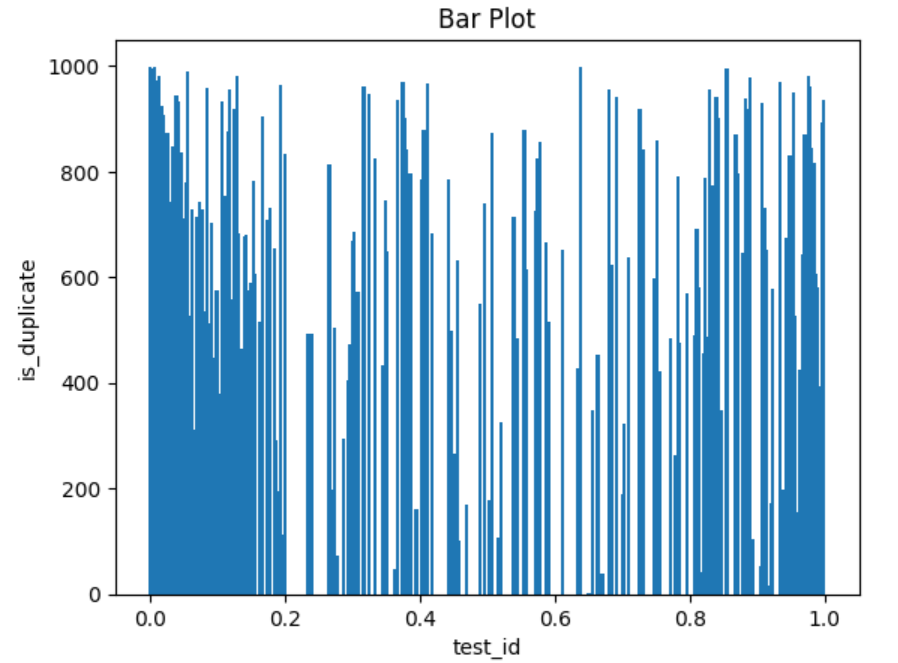
A screenshot of a questionnaire

Description automatically generated

Head 1000 data of the result can show like this.

A diagram of blue dots

Description automatically generated



Reference:

<https://www.geeksforgeeks.org/matplotlib-pyplot-scatter-in-python/>

<https://huggingface.co/bert-base-uncased>

<https://wandb.ai/wandb/common-ml-errors/reports/How-to-Save-and-Load-Models-in-PyTorch--VmlldzozMjg0MTE>

<https://peaceful0907.medium.com/sentence-embedding-by-bert-and-sentence-similarity-759f7beccbf1>

<https://www.kaggle.com/competitions/quora-question-pairs/overview>

<https://www.geeksforgeeks.org/seaborn-barplot-method-in-python/>

<https://stanford.edu/~shervine/blog/pytorch-how-to-generate-data-parallel>