**Question Pair Similarity**

<https://blog.jovian.ai/quora-question-pair-similarity-case-study-d3aac540778f>

<https://medium.datadriveninvestor.com/quora-question-pairs-similarity-a-real-life-nlp-problem-fa06b14f588d>

<https://www.youtube.com/watch?v=1fvQU5yPjFs&list=PLKnIA16_RmvY5eP91BGPa0vXUYmIdtfPQ&ab_channel=CampusX>

[campusx-official/quora-question-pairs: A NLP project to find weather given 2 questions are same are not semantically speaking. (github.com)](https://github.com/campusx-official/quora-question-pairs)

DATASET: [Quora Question Pairs | Kaggle](https://www.kaggle.com/competitions/quora-question-pairs/overview)

* Used ML approach, although can improve using DL techniques, which will be my future work
* Will be an end-to-end project, will create a simple web app and deploy the ML models with Heroku
* Real-life NLP problem, used by many companies like Quora, Stack overflow
* Currently, Quora is facing the issue of Duplicate questions. For instance, consider a pair of questions as follows:
* Question1: What should I do to be a great Geologist?
* Question2: How can I be a good Geologist?
* Notice, the above 2 questions are talking about the same thing (meaning-wise) with the only difference being a different way of asking the same question. Now the challenge for Quora is to eliminate the question pairs that are duplicates of each other (meaning-wise). So that someone does not have to re-write the same answer for the same question asked differently (by using different words but meaning-wise they are the same). This way Quora can enhance the customer experience & reduce the unanswered duplicate questions & thus saving a lot of time for the users.
* Now Quora had organized a Kaggle competition where the participants had to build a machine learning algorithm that can classify whether a given pair of Questions are similar or not.
* These questions asked are very similar with the same intent, but with different wordings and different ways of writing. Hence, we can merge these questions. This can save a lot of time and improves the customer experience.

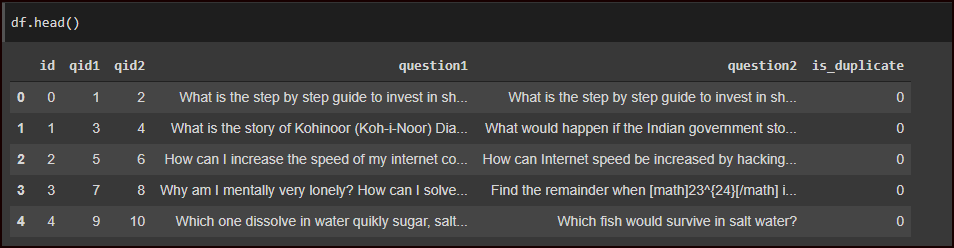
**Problem statement:**

We are tasked with predicting whether a pair of questions that are already asked are duplicates or not.

This could be useful to instantly provide answers to questions that have already been answered.

**Data Overview:**

* Data will be in the file Train.csv.
* Train.csv contains 5 columns : qid1, qid2, question1, question2, is\_duplicate
* Size of Train.csv — 60MB
* Number of rows in Train.csv = 404,290
* ‘is\_duplicate’ is the target label which is 0 for nonsimilar questions and 1 for similar questions.



**Type of Machine Learning problem:**

It is a binary classification problem, for a given pair of questions we need to predict if they are duplicates or not.

The classes are not perfectly balanced, but it’s workable.

* Total number of question pairs for training: 404290
* Question pairs are not Similar (is\_duplicate = 0): 63.08%
* Question pairs are Similar (is\_duplicate = 1): 36.92%
* Total num of Unique Questions are: 537933

**Approach:**

* Will perform EDA, check class imbalance, null values, repetitive qts
* Performance Metric: log loss func
* Will do Text-preprocessing -> removing stop words, punctuation, conv into lower case, etc
* Convert into Vectors using BOW & Word2Vec/Glove(SpaCy) (IR tasks)

Take a step further

* Do Feature Extraction – create new features in order to increase the model accuracy
* ML approach – RF & XGboost algorithms
* Streamlit – simple interface
* Will deploy the models with Heroku