Project Synopsis Information Retrieval 2022

Name of Student: **Sneegdh Krishnna**

Registration No: **199302164**

1. **Project Title:** Identifying Question Pairs Similarity

2. **Overview of Project (8-10 lines):**

   Task is to predict 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. It is a 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?*

  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.

  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)

  Will Take a step further and make it and end-to-end project:

- 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

3. **Any previous work done in same field (if any, mention the detail):**

   The dataset was taken from Kaggle. Quora had hosted a Kaggle Competition where one had to apply advanced techniques to classify whether question pairs are duplicates or not. Doing so will make it easier to find high quality answers to questions resulting in an improved experience for Quora writers, seekers, and readers.

4. Dataset:

i. Dataset Description:

- Train.csv file
- Size of Train.csv — 60MB
- Number of rows in Train.csv = 404,290
- Contains 5 columns : qid1, qid2, question1, question2, is_duplicate
- 'is_duplicate' is the target label which is 0 for non-similar questions and 1 for similar questions

```
df.head()
```

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0

ii. Dataset Features:

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: 537933

iii. Dataset Link: Quora Question Pairs | Kaggle