Overview

Develop machine learning and natural language processing systems to automatically identify when questions with the same intent have been asked multiple times on Quora. Doing so will make it easier to find high-quality answers to questions resulting in an improved experience for Quora writers, seekers, and readers.

Problem Statement

Quora is a place to gain and share knowledge—about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers which empowers people to learn from each other. Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. 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.

More formally, the duplicate detection problem can be defined as follows: given a pair of questions *q1* and *q2,* train a model that learns the function:

$f(q1, q2) \rightarrow 0 \text{ or } 1$

where *1* represents that *q1* and *q2* have the same intent and *0* otherwise.

Course of action:

Requirements:

Dataset is available on http://qim.fs.quoracdn.net/quora_duplicate_questions.tsv released by Quora itself. Our dataset consists of over 400,000 lines of potential question duplicate pairs. Each line contains IDs for each question in the pair, the full text for each question, and a binary value that indicates whether the line truly contains a duplicate pair.

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

Approach to the problem:

We plan to use Siamese LSTM with pre-trained Glove embedding, TF-IDF Vectorizer,

Some of the features we are planning to implement -:

- Similarity measures on the bag of character n-grams (TFIDF reweighted or not) from 1 to 8 grams.
- Edit and sequence matching distances, the percentage of common tokens up to 1, 2, …, 6 when question ends the same or starts the same
- Length of questions, a diff of length
- The number of capital letters, question marks etc…
- Indicators for Question 1/2 starting with "Are", "Can", "How" etc… and all mathematical engineering corresponding