# Project 5 - Is the intent of two questions similar?


## Jason Salazer-Adams


## Overview

There are number of question and answer sites that provide invaluable information for those that are learning a new technical language, i.e. StackOverflow, or those who have general questions about any topic, i.e. Quora. The goal of these sites is to prevent the wheel from being re-invented, and make the next person that much more efficient. However, if you look at any one question, then you will see other users marking questions as duplicates or linking to other questions. This can be very frustrating for the asker and the answerer. The asker spent time crafting some question, because they could not actually find a similar enough question. The answerer is frustrated as they feel they are answering the same question over and over again. In reality, the answerer would rather share their knowledge on a diverse set of topics, instead of the same one.

I am going to attempt to apply ML to help automatically identify whether or not two questions have the same intent.

## Data

I am planning to use a Kaggle dataset provided by Quora. The dataset has 404k pairs of question manually labeled as similar intent or not. These are all real questions posed by users of Quora, and then manually labeled.

## Models

Talking with Roberto, a quick MVP would be converting the questions to a NMF space and performing a similarity metric between the two questions. This can be a good baseline to then build a more complicated model. I was thinking about siamese network architecture as I am comparing two things and want to know if they are similar or not. I am hoping there is a RNN or LSTM text processing network I could use to put into the siamese network architecture, if I get this far.

## Applications

I think this would be a cool input into a question and answer type ofmodel. Could you map questions with the same intent to the same answer? I am also thinking of building a flask app demonstrating how likely two questions are similar. Need to think more how to productize the model.