Problem Statement Formation

Given pairs of questions posted on a Quora, we want to determine whether the intent for each pair is identical. For instance, *"What is the most populous state in the United State?"* and *"Which state in the United States has the most people?"* are asking the same question and users should be guided to a single page.

Context

Quora is a public internet forum where users can post and answer questions on various topics. To avoid publishing duplicate pages that are essentially asking the same question, Quora is looking to utilize machine learning and data science techniques to identify similar questions so as to direct the user to a single source that matches their question.

Criteria for Success

1. Obtain an accuracy score of at least 80% meaning we correctly classify if a pair of questions have the same intent at least 80%. Same goes for correctly identifying pairs that do not ask the same question. The f-1 score metric and confusion matrix will be extremely important.

Scope of Solution Space

1. Verify that probability threshold correctly predicts the value 1 for 'is_duplicate' as 1 indicates similarity
2. How generalizable is our new model? Can we accurately predict on a new data set?
3. Is NLP and neural networks the only available tools?

Constraints

1. Lots of NLP and neural network tools at arsenal
2. Each pair of questions is unique so obtaining a generalizable model can be difficult

Stakeholders

1. Manager
2. Product Managers

Data Sources

1. Kaggle