Task description

You will help us understand how people learn to **simulate the behavior of NLP models.**

DUPLICATE QUESTION DETECTION

The NLP model in question predicts whether two questions are duplicates of each other - that is, whether they should point to the same answer in a website like Quora.The label are thus (*Non-duplicate* / *Duplicate*).

THE CONTEXT

In each round, you will see a reference example like below, with the model's prediction on it (below, the model makes a `Correct` **prediction**).

`Old Q1` Why is the sky blue ?
`Old Q2` Why is that the sky is so blue ?
`Model predicts` *Duplicate (98.5% confident)*
`Model correct?` **`Correct`**

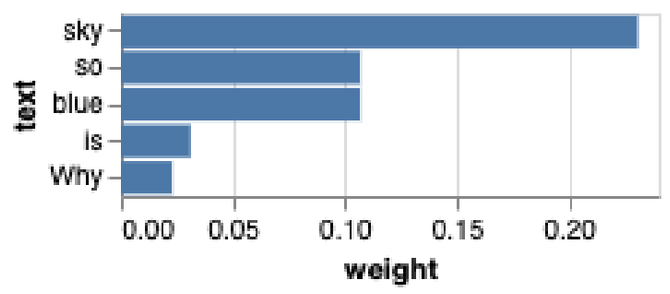
Additional Clues about the model's behavior

FEATURE IMPORTANCE

To help you understand what is driving the model's prediction, we keep **Q1** fixed and highlight on **Q2** words that are important for the model's prediction, in blue (we also show a bar chart with the five most important words) .

These values are computed by a standard black-box explanation technique (SHAP), which **masks different groups of words in Q2** and summarizes how predictions change as a result.

`Old Q1` Why is the sky blue ?
`Old Q2` Why is that the sky is so blue ?
`Model predicts` *Duplicate (98.5% confident)*
`Model correct?` **`Correct`**



ASK THE MODEL QUESTIONS!

You can also **make small changes to Q2** and see the resulting model predictions, in order to get a better understanding of how the model behaves around the reference example. You can ask up to **10** additional questions per round to learn more about the model. For example, you may want to ask the following question:

`Old Q1` Why is the sky blue ?
`New Q2` Why is that the sky is so • dark ?
`Model predicts` *Non-duplicate (99.6% confident)*

YOUR TASK

After seeing *the reference example, feature importances, and asking your own questions of the model*, we will ask you to try to **guess how the model would predict several variations of Question 2** (`New Q2`).

Beware that **the model is not perfect, so it may make mistakes** (you should try to simulate the model to the best of your ability). Below is an example of a variation of `Q2` we might ask you to label:

`Old Q1` Why is the sky blue ?
`New Q2` Why is that the sky is so • white ?
`Model will predict` ○ *Non-duplicate* ◉ *Duplicate*

As you see, the model may be incorrect, so please learn about **the model's behavior** carefully through the **Additional Clue**.

Procedure

You will first go through a **1-round training phrase** to help you get familiar with the task. Then, you will complete **20 rounds** of labelings.