Udacity Machine Learning Nanodegree Capstone

Quora Duplicate Question Detection

Raahul Seshadri

April, 2017

Contents

1 Definition

1.1 Project Overview

Websites like Quora[1], StackExchange[2] network (includes the popular programming website StackOverflow) allow the users to post questions, and the entire community can answer those questions. Ideally, every unique question should be present just once in the system, so that every answer to those questions are present at once place.

However, users are prone to posting duplicate questions, because not all of them check if the question they're asking has already been asked by someone else. This necessitates having an automated duplicate question checker that can check if a question is a duplicate. Whenever the user tries to post a new question, the system can suggest an existing one for perusal.

This project was inspired by Quora's Kaggle challenge. A dataset of question pairs, manually tagged as duplicate or note, has been provided by Quora to train on.

1.2 Problem Statement

The problem can be stated as follows:

Given a question pair, detect if they are duplicate or not.

Quora provides a dataset

[1] https://www.quora.com

[2] https://stackexchange.com