Udacity Machine Learning Nanodegree Capstone

Quora Duplicate Question Detection

Raahul Seshadri

April, 2017

Contents

1 Definition **2**
1.1 Project Overview 2
1.2 Problem Statement 2
1.3 Metrics 2

2 Analysis **3**
2.1 Data Exploration 3

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[3]. 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.

The steps to achieving our duplicate detector are as follows:

1. Download and pre-process the Quora training dataset
2. Split the Quora training dataset into 90% training and 10% test set.
3. In the 90% training set, further 10% will be used as a validation set.
4. Extract usable features from the dataset.
5. Train a binary classifier to differentiate between duplicates/non-duplicates.
6. Provide a command-line interface so that users can check for duplicates from Quora's test set, or provide his own questions.

1.3 Metrics

Being a classification problem, the metric is going to be accuracy:

$$\text{accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{total samples}}$$

This accuracy will be measured on the test set (10% of the total dataset).

[1] https://www.quora.com
[2] https://stackexchange.com
[3] https://www.kaggle.com/c/quora-question-pairs

2 Analysis

2.1 Data Exploration

Let's look at how the Quora dataset is structured:

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0
5	5	11	12	Astrology: I am a Capricorn Sun Cap moon and c...	I'm a triple Capricorn (Sun, Moon and ascendan...	1
6	6	13	14	Should I buy tiago?	What keeps childern active and far from phone ...	0
7	7	15	16	How can I be a good geologist?	What should I do to be a great geologist?	1
8	8	17	18	When do you use シ instead of し?	When do you use "&" instead of "and"?	0
9	9	19	20	Motorola (company): Can I hack my Charter Moto...	How do I hack Motorola DCX3400 for free internet?	0
10	10	21	22	Method to find separation of slits using fresn...	What are some of the things technicians can te...	0
11	11	23	24	How do I read and find my YouTube comments?	How can I see all my Youtube comments?	1
12	12	25	26	What can make Physics easy to learn?	How can you make physics easy to learn?	1
13	13	27	28	What was your first sexual experience like?	What was your first sexual experience?	1