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[3]) 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[4]. A dataset of question pairs, manually tagged as duplicate or not by human reviewers, has been provided[5] by Quora to train on.

1.2 Problem Statement

The problem can be stated as follows:

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

For example, given the following two questions:

Q1: What is the average salary in India?
Q2: What is the average salary in the United States?

The system could flag them as either duplicates or not-duplicates. 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] https://www.quora.com
[2] https://stackexchange.com
[3] https://stackoverflow.com
[4] https://www.kaggle.com/c/quora-question-pairs
[5] https://www.kaggle.com/c/quora-question-pairs

1.3 Metrics

Being a binary 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).

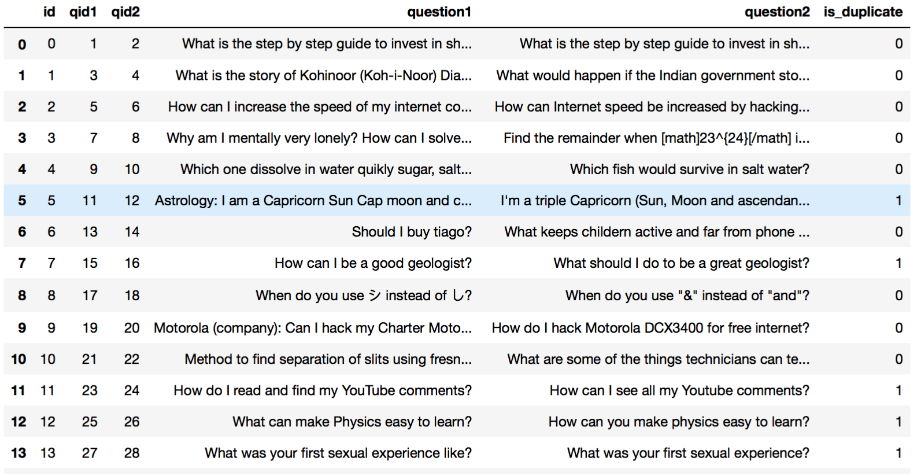
1. **False negatives**, where a duplicate pair is not detected as such, will lead to degraded user experiences. However, it is also rather harmless, since no information is lost. There is still an opportunity to manually flag them as duplicates, or build such a feature on the platform that is using the duplicate detector.

2. **False positives**, where a question pair that's not a duplicate, but is marked as such, is more dangerous. It can lead to very degraded user experience, since it directly hinders the user trying to perform an action. It is better for the system to err on the side of false negatives than false positives.

2 Analysis

2.1 Data Exploration

Below is the head (first 20 entries) of the Quora dataset:



	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

The columns correspond to the following features:

1. **id:** ID to identify the question pair uniquely

2. **qid1:** ID of the first question in the question pair

3. **qid2:** ID of the second question in the question pair

4. **question1:** Text of the first question in the question pair

5. **question2:** Text of the second question in the question pair

6. **is_duplicate:** Quora reviewer's decision on whether the question pair is duplicate (1) or not (0)

2.1.1 Why do we have question IDs for each pair?

There are two columns of interest, ”qid1” and ”qid2” that are of interest. The algorithm that we're going to design will only tell us if a given pair of question is a duplicate. However, given a question, we'd also like to find out all the duplicates of it in the system.

For example, if question ”1” is a duplicate of question ”2”, and question ”2” is a duplicate of ”3”, then a system should also know that ”1” and ”3” are duplicates. Data structures like ”union find” allow us to do that. However, this is not in scope of the algorithm that we'll design, but the responsibility of the higher system that will use this duplicate detector.

Which is why the question IDs won't be used as a feature for the duplicate detector, but is still important for the system using the duplicate detector.

2.1.2 Summary statistics

Below are some basic summary statistics:

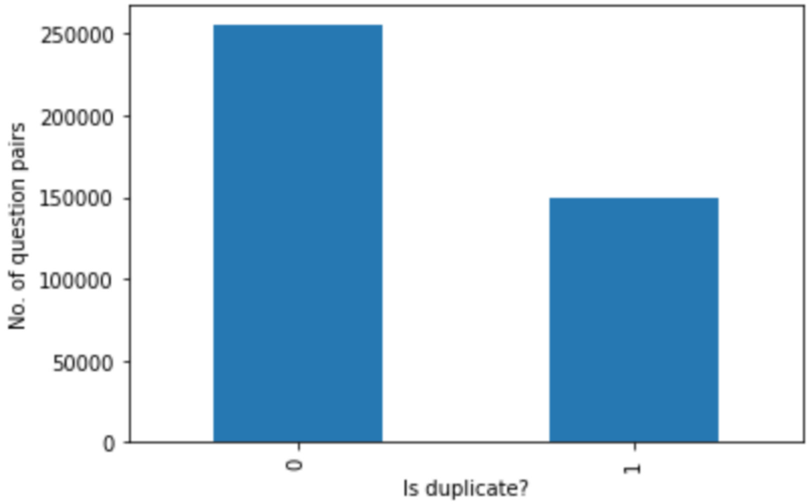
1. **Total entries:** 4,04,290
2. **Total positive entries:** 1,49,263
3. **Total negative entries:** 2,55,027
4. **Percent positive entries:** 36.91%
5. **Percent negative entries:** 63.08%

We have a sample with more negative than positive examples.

2.2 Exploratory Visualization

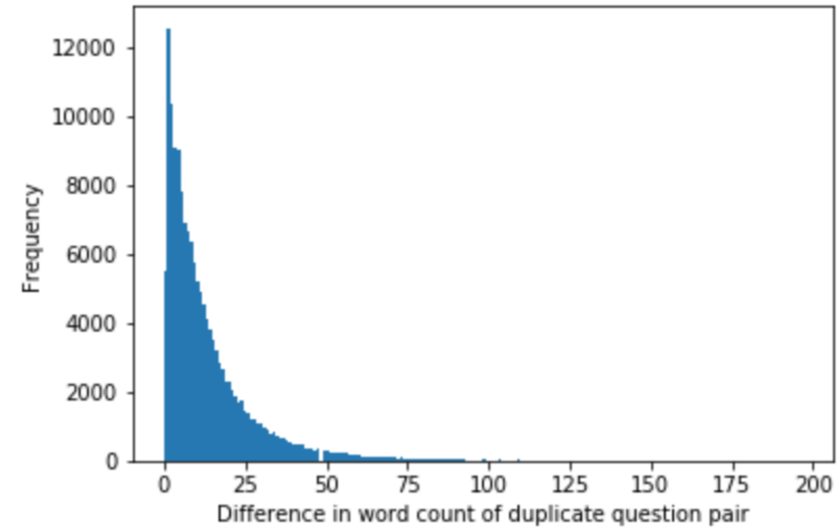
2.2.1 Summary visualization

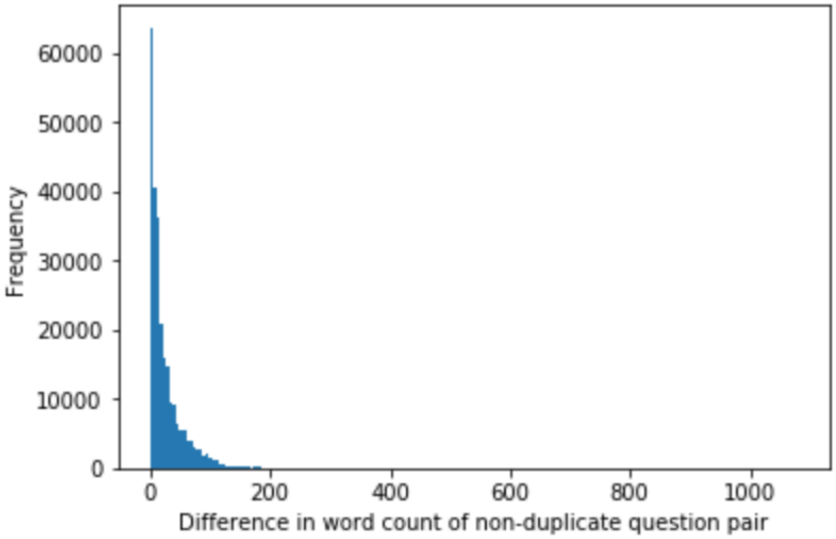
Let's look at the class distribution (duplicate or not). This is the same from our summary statistics.



2.2.2 Word count differences

Let's look at if word count has any bearing on whether a question is marked as duplicate or not, or can we discern any important facts/patterns.





1. For duplicates, the word count difference tapers smoothly

2. For non-duplicates, there is a sharp decrease in the 2nd bin

3. There are question pairs that have word count difference 75 and above, and are still duplicates. So simple bag of words model will not work. *This is the most important conclusion.*

2.2.3 Jaccard index

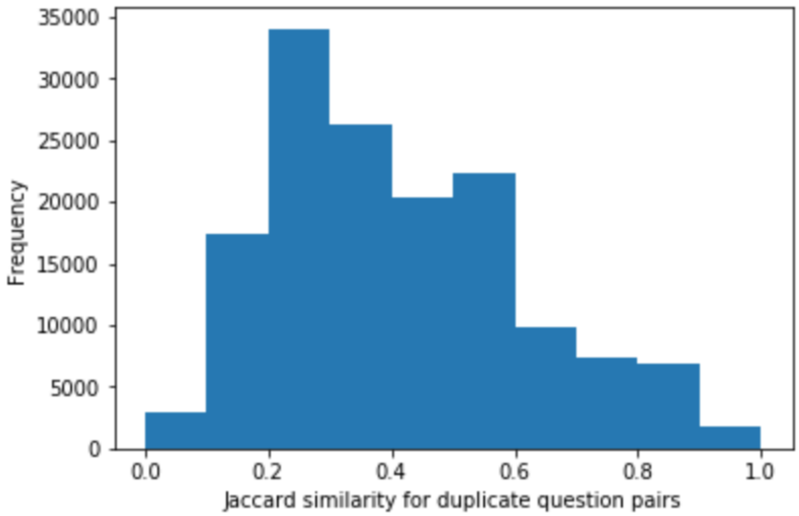
To explore a baseline model, let's look at the bag-of-words similarity of duplicate and non-duplicate question pairs using the Jaccard Index[6]. The relevant formula is:

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

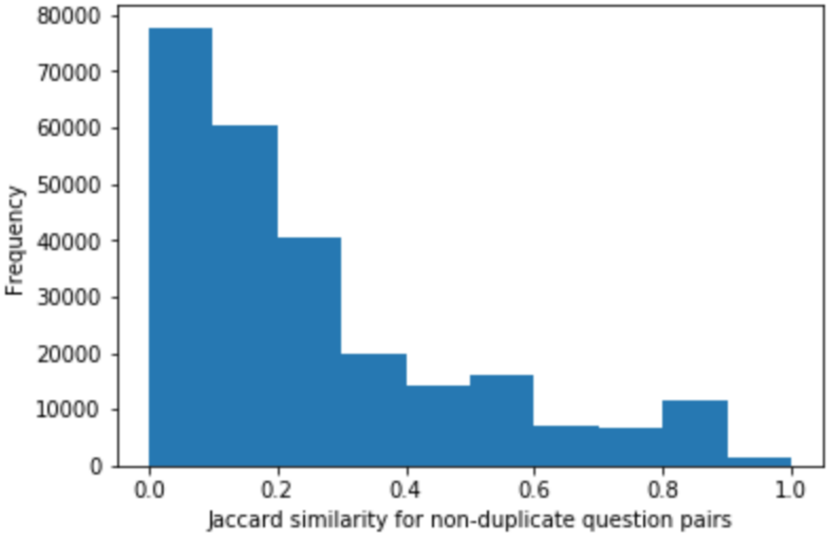
In our case, A and B refer to individual words in the question pairs. So, the formula basically is:

$$J(A,B) = \frac{\text{No. of common words in question1 and question2}}{\text{No. of unique words in question1 and question2 combined}}$$

Let's look at the Jaccard similarity of duplicate and non-duplicate question pairs separately.



[6] https://en.wikipedia.org/wiki/Jaccard_index


80000
70000
60000
50000
40000
30000
20000
10000
0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Jaccard similarity for non-duplicate question pairs