Quora Question Pairs Dataset Analysis

Xinchen Pan
Department of Statistics
Univeristy of Illinois at Urbana-Champaign
pan31@illinois.edu

May 8, 2017

Abstract

For this project, we conduct a complete analysis on the Quora Duplicated Questions dataset using different methods in text mining including topic modeling, document clustering and semantic analysis. We build several prediction models like logistic regression, random forest and gradient boosting classifier to compare the accuracy of duplicated question predictions. Besides that, we also present some visualization works like bar plots and word clouds to show some aspects of the dataset.

1 Introduction

1.1 Background

Quora is a website for people around the world to ask and answer questions in English. It is not surprising that there is a high chance of duplicated questions asked by users. Thus it is necessary to have a model which can detect whether a certain problem has already been asked. At the beginning of 2017, Quora posted its first dataset which is called question pairs dataset at `https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs`. We used this dataset to achieve the following two main goals. One is to do the topic modeling & document clustering to see if any hidden topics can be found from the questions. Another is to build prediction models to predict whether two questions are duplicated.

1.2 Dataset Description

The dataset is in a structure as shown in Table 1 There are a total of 6 columns: `id`, `qid1`, `qid2`, `question1`, `question2` and `is_duplicate`. The first three columns are not helpful for our analysis as they are all index information. Thus we only used last three columns for our

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers ?	What is a least natural number?	0
1518	3038	3039	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3372	6542	6543	How do you start a bakery ?	How can one start a bakery business?	1
3272	6542	6543	Should I learn Java or Python first?	If I had to choose between learning Java or Python, what should I choose to learn first?	1

Table 1: An overview of the dataset

analysis. `question1` and `question2` are the contents of the questions. `is_duplicate` tells whether this question pair is duplicated.

Table 2 gives a summary of the dataset. The longest question in the dataset has 237 words while the shortest question has only 1 word in length.

	Question 1	Question2
Number of question pairs	404351	404351
Average number of words	10.944184	11.184303
Standard deviation of number of words	5.431847	6.311142
Max number of Words	125	237
Min number of words	1	1

Table 2: Summary Statistics of the dataset

1.3 Stop words

Stop words are a set of common words that account for a large percentage in the language but do not have much meaning. For example, the most common stop words in English are "the", "am", "is" etc. It is usually necessary to remove stop words in natural language processing but it is not absolute.

Since our text are all questions, we will expect to have a large percentage of "What is", "How is", "Why". Removing these kind of question words might be useful in predicting duplicated questions.

1.4 N-grams

N-grams are a contiguous sequence of words. By finding out most frequent n-grams, we are able to know what kind of questions are asked most in these 0.4 million question pairs.

In order to build n-grams, we performed the following procedures. Firstly, we converted all questions strings to lower cases. Then we tokenized the questions. Next we removed stop words and punctuations. Then we used `Countvectorizer` from `scikit-learn` library in python to count the frequency of each word. Here "natural" and "numbers" are both unigrams, "natural numbers" is a bigram.

"What are natural numbers?"
⇓
"what are natural numbers?"
⇓
"What", "are', "natural", "numbers", "?"
⇓
"natural", "numbers", "?"
⇓
"natural", "numbers"

Because we have two columns of questions, I combined both of the questions if they are not duplicated questions to make n-grams.

An n-gram of size 1 is called "unigram", size 2 is called "bigram", size 3 is called "trigram", so on and so forth.

Figure 1 is the most frequent 15 unigrams. As we can see, "best" has the highest frequency. It makes perfect sense because when people ask recommendations, they will definitely want the "best" option. The third most frequent unigram is India. According to Alexa, a California-based company that provides commercial web traffic data and analytics, about 16% of Quora users are from India.(1)

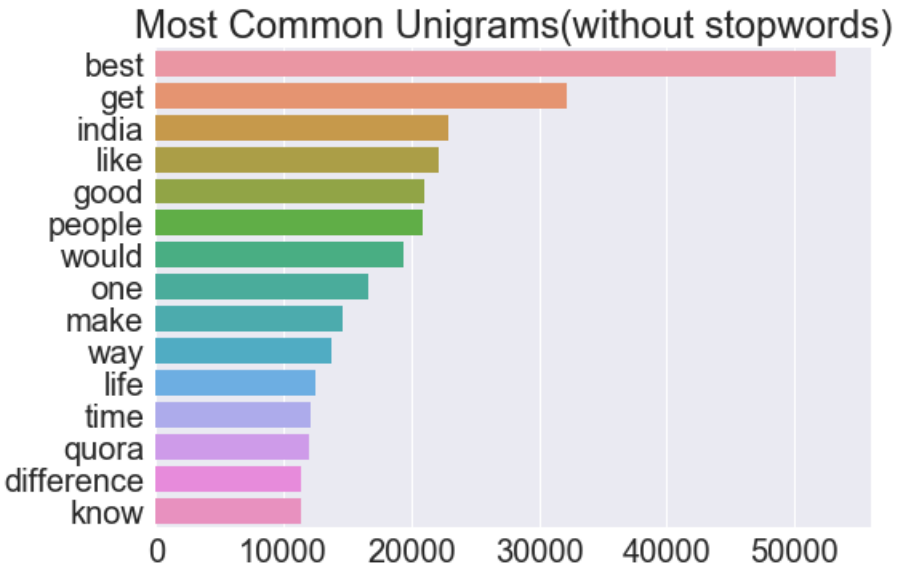

Figure 1: Top 15 frequent unigrams

Figure 2 is the most frequent 15 bigrams. It is not surprising that the most frequent bigrams is "best way". We already know that the we have the most "best" in unigrams.

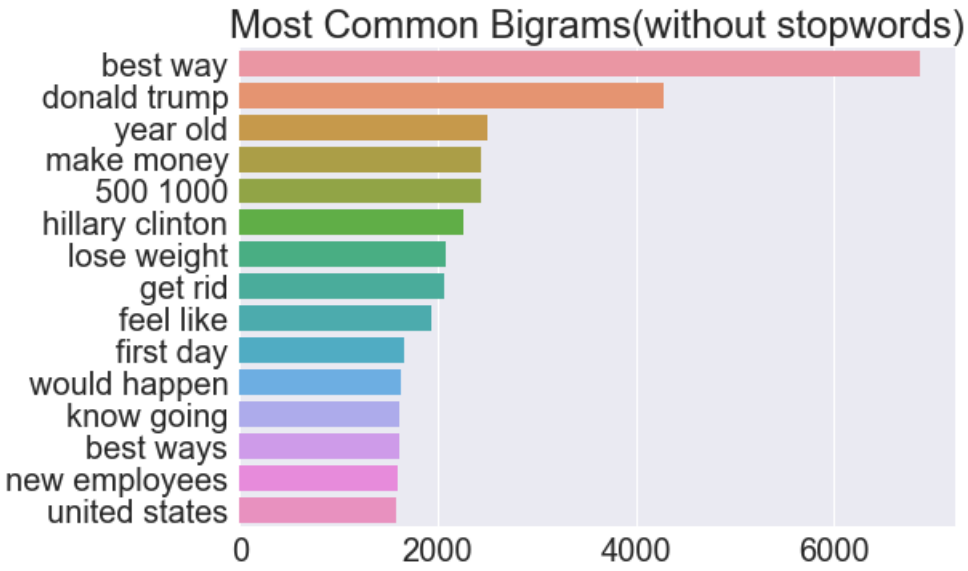

Figure 2: Top 15 frequent bigrams

Besides asking about "best book", "best method", it is more general to ask the "best way" to learn/do sth. The second most common bigrams is "donald trump" and the sixth most common bigrams is "hilary clinton". We do not need much explanation for this since the 2016 American President election has conquered the first page of the news for a very long time.

For a comparison, we made a bar plot for the most common top 15 trigrams without removing stop words as shown in Figure 3. We can clearly see that almost all trigrams are questions words.

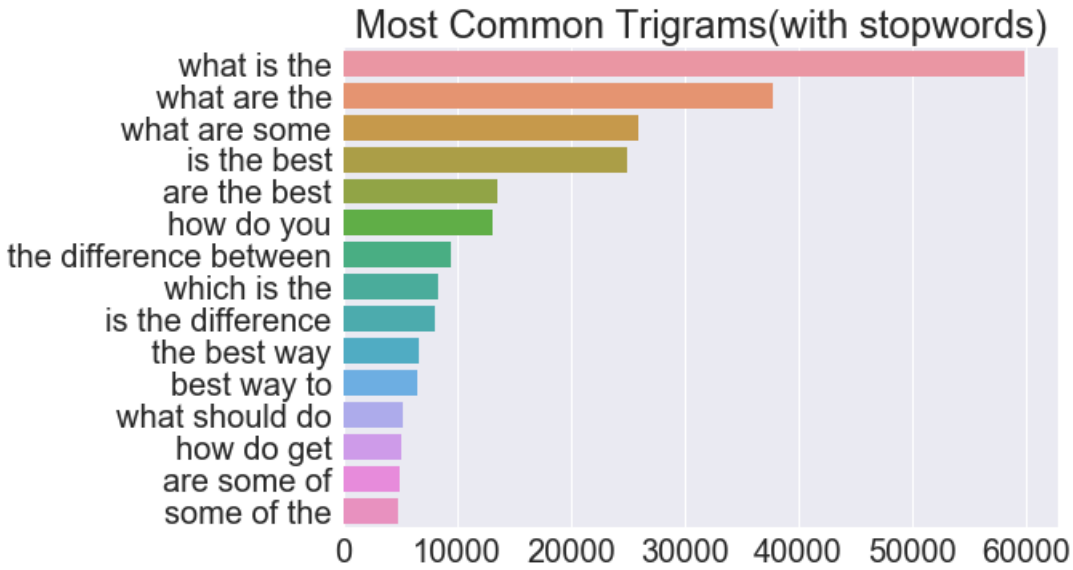

Figure 3: Top 15 frequent bigrams

1.5 Word Cloud

Word cloud is a visual representation of text data. By building word clouds, we are able to find out the most common words in the text as the size of the words indicates the frequency. Figure 4 is a masked word cloud. Words like "best way", "difference", "india", "donald trump" can all be easily found.

Figure 4: Top 15 frequent bigrams

2 Topic Modeling

When we first register in Quora, we can choose what categories of questions we are interested in, like "history", "philosophy" or "math". Thus, it is reasonable for us to believe that the 0.4 million question pairs are in a set of topics. We can use some statistical methods to discover these topics.

Generally speaking, a topic model is a statistical model which can be used to discover the "topics" within a collection of documents.

We followed the following steps to do topic modeling.

Lower case ⇒ Tokenization ⇒ Stemming ⇒ Lemmatization ⇒ Compute term-frequency

After all these steps, we will get a term-frequency matrix which will be used for the following analysis.

2.1 Tf and Tf-idf

The full name for tf is term-frequency. It is simply a counter vector counting number of distinct words for every document. In our case, each question is a document. All 0.4 million questions form a very large dictionary vector. Then every distinct word will be 1 if that word is in the corresponding document in the dictionary vector. All these counter vectors form the term-frequency matrix.

Unlike term-frequency, term frequency-inverse document frequency is used to reflect how important a word is to the document. Some words may be very important but they have a low term frequency.

Tf and tf-idf both have advantages and limitations. For my analysis, I used tf for topic modeling and document clustering.

2.2 Non-Negative Matrix Factorization

Non-negative matrix factorization is a method to do topic modeling. Given a matrix V of $m \times n$ dimension in which every element of matrix V is greater or equal than 0. By doing NMF, we are seeking a matrix W and H with a dimension of $m \times k$ and $k \times n$. k is determined by us. Matrix W and H are determined by minimizing th following

$$||V - WH||^2$$

Figure 5 is a illustration of non-negtaive matrix factorization.(2)

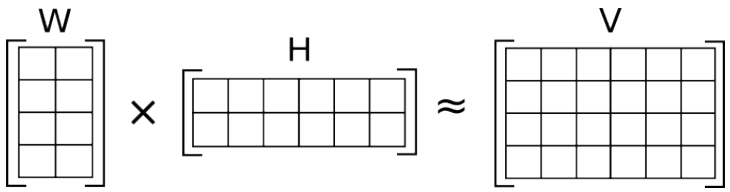

Figure 5: Non-Negative Matrix Factorization

For topic modeling, V is the term document frequency or term frequency inverse document frequency matrix. k is the number of topics we choose. For this project, we chose 10 topics and got the following result in Figure 6

Checking the first topic, we may guess that this topic is about recommendations. It is common for people for ask recommendations for books, movies. For second topic, it is about politics as we find "donald" and "trump" in it. Topic 5 is about the why Indian government demonetized 500, 1000 rupee bank notes in 2016.

We can see that it is not necessary that we can understand what the topic represents from the most frequent words, for example topic 9.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
best	people	india	like	note	good	make	quora	know	way
book	think	pakistan	feel	500	book	money	question	new	learn
movie	know	war	girl	1000	time	online	answer	thing	weight
place	trump	country	look	rupee	job	earn	ask	day	lose
learn	ask	trump	work	r	bad	trump	google	year	easiest
laptop	donald	company	company	indian	idea	youtube	improvement	life	english
buy	believe	job	guy	black	position	friend	asked	time	money
website	world	donald	culture	money	differ	video	easily	going	improve
site	hate	u	woman	banning	department	month	needing	employee	suicide
2016	use	president	different	government	balance	black	answered	want	online

Figure 6: Topics generated non-matrix factorization

2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation is another method for topic modeling. It is a generative probabilistic model which is suitable for text data.

We got the following result in Figure 7 We have some similar topics like NMF and we also have topics that cannot tell what they are about.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
best	life	quora	difference	indian	best	trump	use	people	day
india	phone	like	time	work	good	mean	note	know	want
make	account	question	love	world	way	new	person	think	girl
money	long	start	possible	u	thing	feel	500	best	india
year	google	business	woman	country	learn	student	1000	2016	really
book	say	used	weight	number	better	donald	need	website	friend
movie	help	good	sex	change	company	using	different	computer	war
online	android	lose	stop	like	language	university	black	game	job
engineering	mobile	answer	2	card	college	example	compare	exam	place
way	social	increase	win	site	like	president	free	learning	state

Figure 7: Topics generated by Latent Dirichlet Allocation

3 Document Clustering

3.1 K-means Clustering

Similar to topic modeling, document clustering is another way of clustering documents. A wide used algorithm for document clustering is K-means clustering.

The objective of K-means clustering is to find:

$$\arg\min_{\mathrm{S}} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 = \arg\min_{\mathrm{S}} \sum_{i=1}^{k} |S_i| \mathrm{Var} S_i$$

Where S is our clusters and μ_i is the mean of points for S_i. k is the number of clusters.

We used K-means algorithm to generate 10 clusters and find out the top 10 common words in each cluster following the same procedure as topic modeling.

Checking Figure 8, there are basically no significant differences in topics between the above two methods. But we find that Topic 3 clustered really well because all the words are cosmetology related .

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
people	start	rid	best	mean	like	difference	good	note	india
quora	business	acne	way	say	feel	make	book	500	best
question	learning	fat	book	girl	look	quora	way	1000	pakistan
think	want	way	learn	dream	girl	life	make	rupee	war
know	best	belly	movie	guy	culture	time	time	r	country
ask	preparation	face	place	symbol	work	way	position	money	company
like	programming	pimple	online	word	company	year	department	black	job
google	ia	scar	website	love	guy	know	employee	banning	indian
easily	preparing	addiction	laptop	person	woman	money	balance	indian	available
answer	india	hair	site	phrase	know	thing	differ	ban	spotify

Figure 8: Topics generated by K-means algorithm

4 Duplicated Questions Prediction

An important part of this project is to do predict whether the questions in `questions1` and `questions2` are duplicated.

In order to build the model, we processed the data in these steps just like before.

1. Convert the questions to lower case
2. Tokenize the questions
3. Remove stop words
4. Remove Punctuations
5. Stemming
6. Lemmatization

Stemming is the process to convert the word to its root form. E.g cats" ⇒ "cat". Lemmatization is the process to convert the inflected word to its original form. E.g, better ⇒ good.

After all these steps, we want to generate some features from the processed dataset for the purpose of predicting.

Notice that the percentage of not duplicated questions in our dataset is about 63%. Thus our model needs to have a higher accuracy than this number.

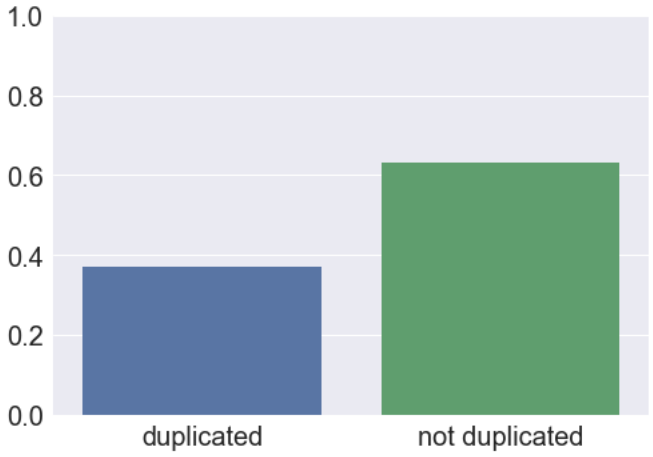

Figure 9: Percentage of Duplicated Questions

The features we made are in Table 3

Features	Description
WordLen1	number of words in question 1
WordLen2	number of words in question 2
StrDiff1	string length of question 1
StrDiff2	string length of question 2
Similarity	average similarity between question 1 and question 2
WordShare	number of common words
WeighShare	weight difference between question 1 and question 2

Table 3: Features description

The reason for having `WordLen1` and `WordLen2` is because the left words after processing contain the most important information of the questions. Thus the number of words will tend to measure how much useful information left for each question. `StrDiff1` and `StrDiff2` are used to measure the original information. `Similarity` variable is the average similarity between words from question 1 and question 2. We used `Word2Vec` from `gensim` to calculate it. `Wordshare` is the number of common words in both questions. `WeightShare` is the weight difference between the term frequency inverse document frequency of the words in two questions.

4.1 Results

Table 4 is our results. We find that gradient boosting classifier using XGBoost library can give us best result.

Methods	Test Accuracy
Logistic Regression	67.5%
Random Forest	69.7%
Gradient Boosting Classifer	70.5%
Gradient Boosting Classifer(Using xgboost library)	70.9%

Table 4: Accuracy of different method

5 Conclusion

From the analysis above, we can get a few conclusions. Most of the questions Quora users like asking are about recommendations, differences and recent hot topics. For example, best way of learning something, difference the 2016 presidential election between Hilary Clinton and Donald Trump, 2016 rupee demonetization in India.

For prediction, checking a recent paper in `arXiv:1702.03814 [cs.AI]`, they are able to achieve a 88% accuracy by using a bilateral multi-perspective matching models. Some other models like Siamese-CNN, Multi-Perspective-CNN, Siamese-LSTM models can achieve accuracy over 80%. Thus still more improvements needed in order to increase accuracy for my model. More evolved word embedding methods need to be examined and data could to be further cleaned.

6 Reference

1. Quora. (2017, April 27). In Wikipedia, The Free Encyclopedia. Retrieved 02:30, April 29, 2017, from https://en.wikipedia.org/w/index.php?title=Quora&oldid=777529144

2. Non-negative matrix factorization. (2017, April 22). In Wikipedia, The Free Encyclopedia. Retrieved 02:29, April 29, 2017, from https://en.wikipedia.org/w/index.php?title=Non_negativex_factorization&oldid=776644650