STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Identifying Duplicate Quora Question Pairs

Final Project: CS 513 KDD

By:
Vaibhav Prasad Desai
Vishnu Pillai
Piyush Bhattad

AGENDA

- Introduction
- Dataset Details
- Feature Engineering
- Machine Learning
- Results
- Conclusion

Introduction

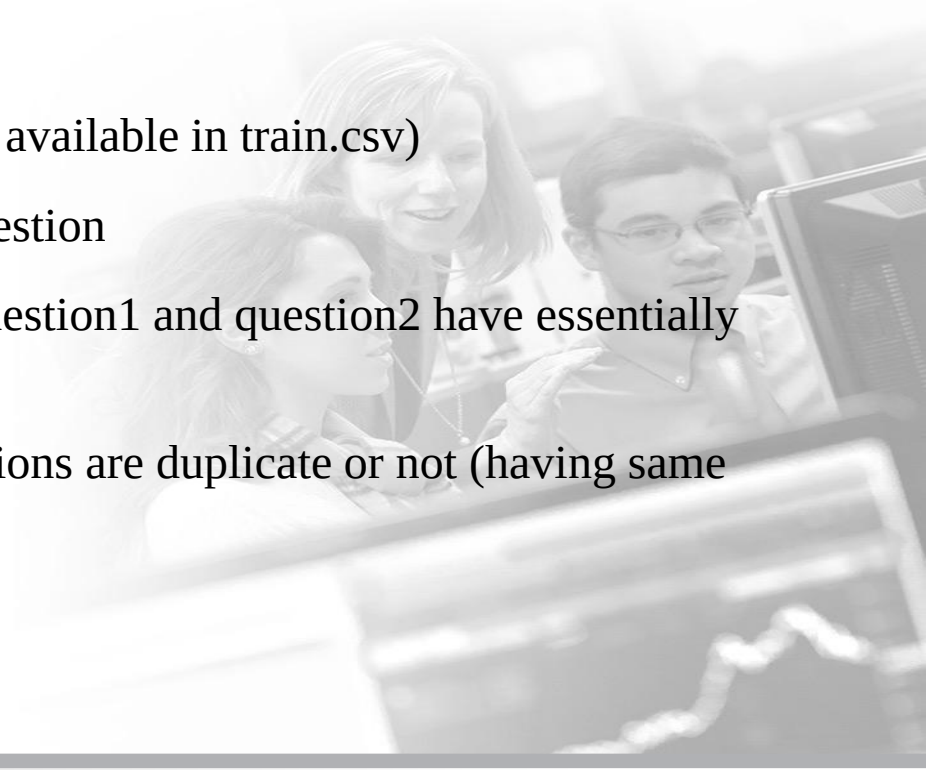
- Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question.
- It is important that we ensure each unique question exists on Quora only once. Writers shouldn't have to write the same answer to multiple versions of the same question, and readers should be able to find a single canonical page with the question they're looking for.
- Develop machine learning and natural language processing system to automatically identify when questions with the same intent have been asked multiple times.

DATA

train.csv

- id - the id of a training set question pair
- qid1, qid2 - unique ids of each question (only available in train.csv)
- question1, question2 - the full text of each question
- is_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

- The Objective is to predict whether the Questions are duplicate or not (having same intent or not)



DATA

id	qid1	qid2	question1	question2	is_duplicate
0	1	2	What is the step by step guide to invest in shar	What is the step by step guide to invest in share market?	0
1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia	What would happen if the Indian government stole the Kohinoor (Koh-i-Noor)	0
2	5	6	How can I increase the speed of my internet cc	How can Internet speed be increased by hacking through DNS?	0
3	7	8	Why am I mentally very lonely? How can I solv	Find the remainder when [math]23^{24}[/math] is divided by 24,23?	0
4	9	10	Which one dissolve in water quikly sugar, salt,	Which fish would survive in salt water?	0
5	11	12	Astrology: I am a Capricorn Sun Cap moon and	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this	1
6	13	14	Should I buy tiago?	What keeps childern active and far from phone and video games?	0
7	15	16	How can I be a good geologist?	What should I do to be a great geologist?	1
8	17	18	When do you use ã,· instead of ã—?	When do you use "&" instead of "and"?	0
9	19	20	Motorola (company): Can I hack my Charter M	How do I hack Motorola DCX3400 for free internet?	0
10	21	22	Method to find separation of slits using fresnel	What are some of the things technicians can tell about the durability and relia	0
11	23	24	How do I read and find my YouTube comment	How can I see all my Youtube comments?	1
12	25	26	What can make Physics easy to learn?	How can you make physics easy to learn?	1
13	27	28	What was your first sexual experience like?	What was your first sexual experience?	1
14	29	30	What are the laws to change your status from	What are the laws to change your status from a student visa to a green card ir	0
15	31	32	What would a Trump presidency mean for cur	How will a Trump presidency affect the students presently in US or planning to	1
16	33	34	What does manipulation mean?	What does manipulation means?	1

Preprocessing

question1	question2
.	Why is Cornell's endowment the lowest in the Ivy League?
?	Why should one not work at Google?
deleted	Which website will be suitable for downloading eBooks and lectures?
?	What is the Gmail tech support help phone number?
deleted	Which are some best websites for downloading newly published books/eBook
HH	What is hh?
What?	What should Indians do if Donald Trump becomes President?
deleted	What kind of questions on Quora aren't OK? What is Quora's policy on questic
deleted	What is a website where I can download eBooks legally?
Na	How do I activate Reliance Jio 4G?
I'm	I am a 39 year old single woman. Should I have a baby alone?
grammar	What is grammar?
How long?	How long does a New Zealand citizen immigrant have to wait to apply for his p
What?	What is the average processing time for a spouse sponsorship in Canada after
lol ?	What is League of Legends?
Is?	Does mother/son incest happen in India?
Deleted.	How do I send notification from server to Android without using GCM?
How I am?	Why does red flowers not putting on the Shiv linga?

- Removing Strings of length<9
- Stopwords removal (selective Words like ‘the’, ‘a’, ‘an’.)
- Building Bag-of-Words(Tokenaization)

FEATURE ENGINEERING

- Term Frequency
- Inverse document frequency
- Dot product of above TFIDF
- Vectorization of words
- Summation and average of vectors for each sentence
- Representing each sentence as vector

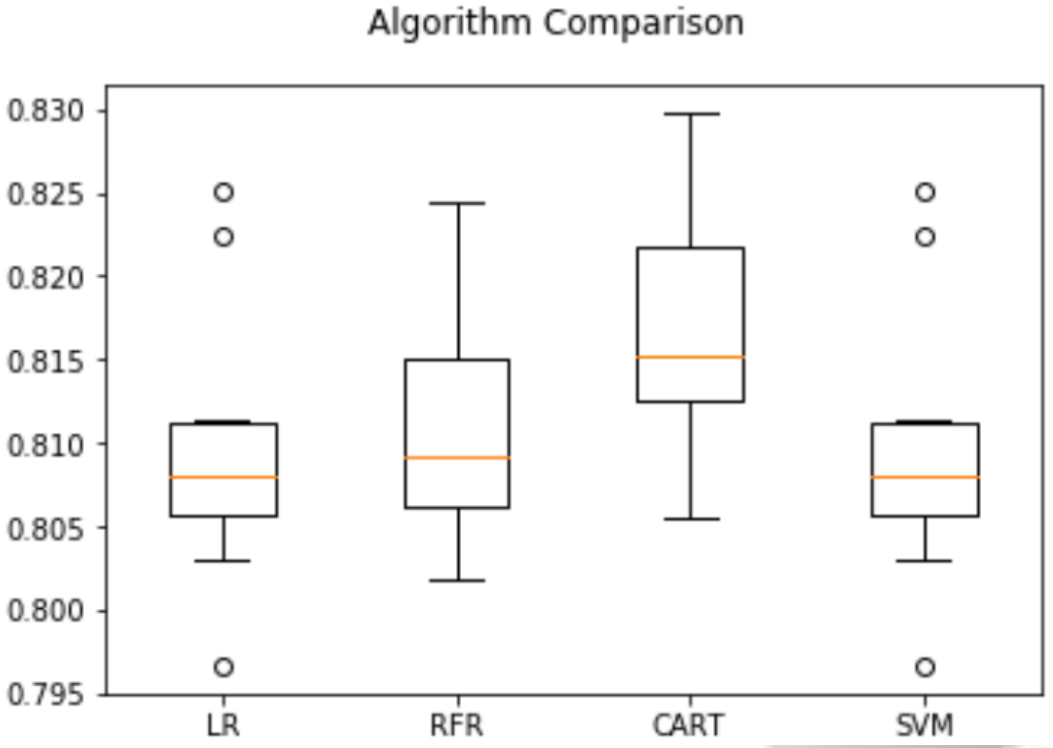
Main features:

Cosine Similarity

Euclidean distance

Machine Learning & Accuracy

1. Logistic Regression
2. Random Forest
3. CART
4. SVM



Improvisation

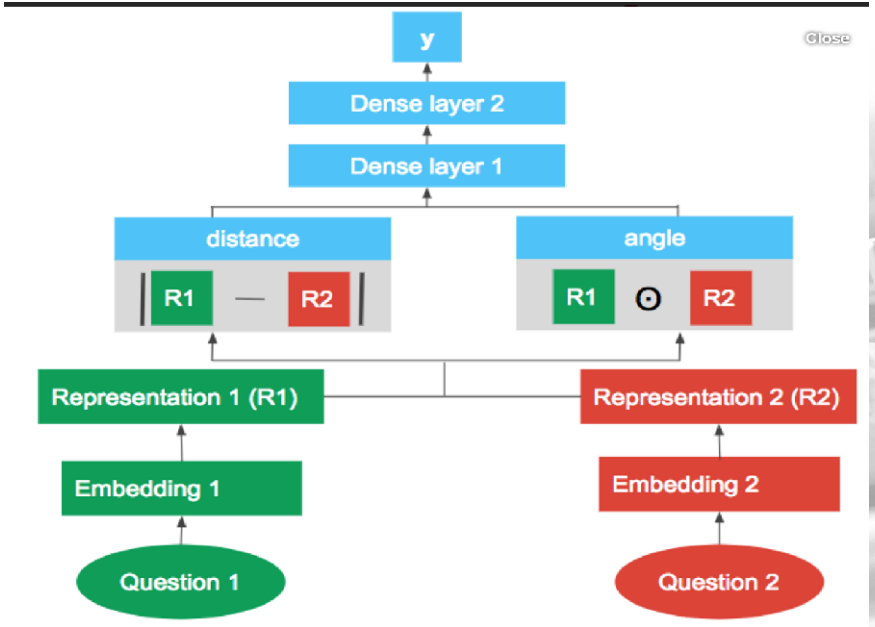
Work with Neural networks using LTSM model

Use Gensim Doc2vec
(based on word2vec)
For vectorization of questions
Further used for deriving

- Cosine similarity.
- Distance function

Alongwith:

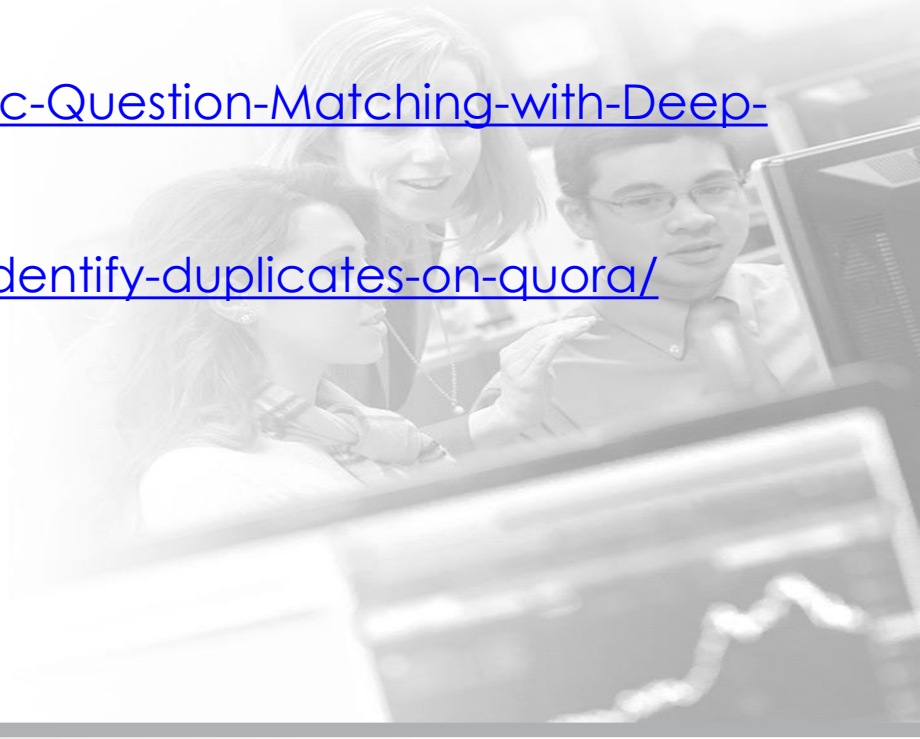
- Common word share(weights)
- Pos Tagging



References:

https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning

http://www.janagrc.com/dupe-snoop-identify-duplicates-on-quora/



Thank You!