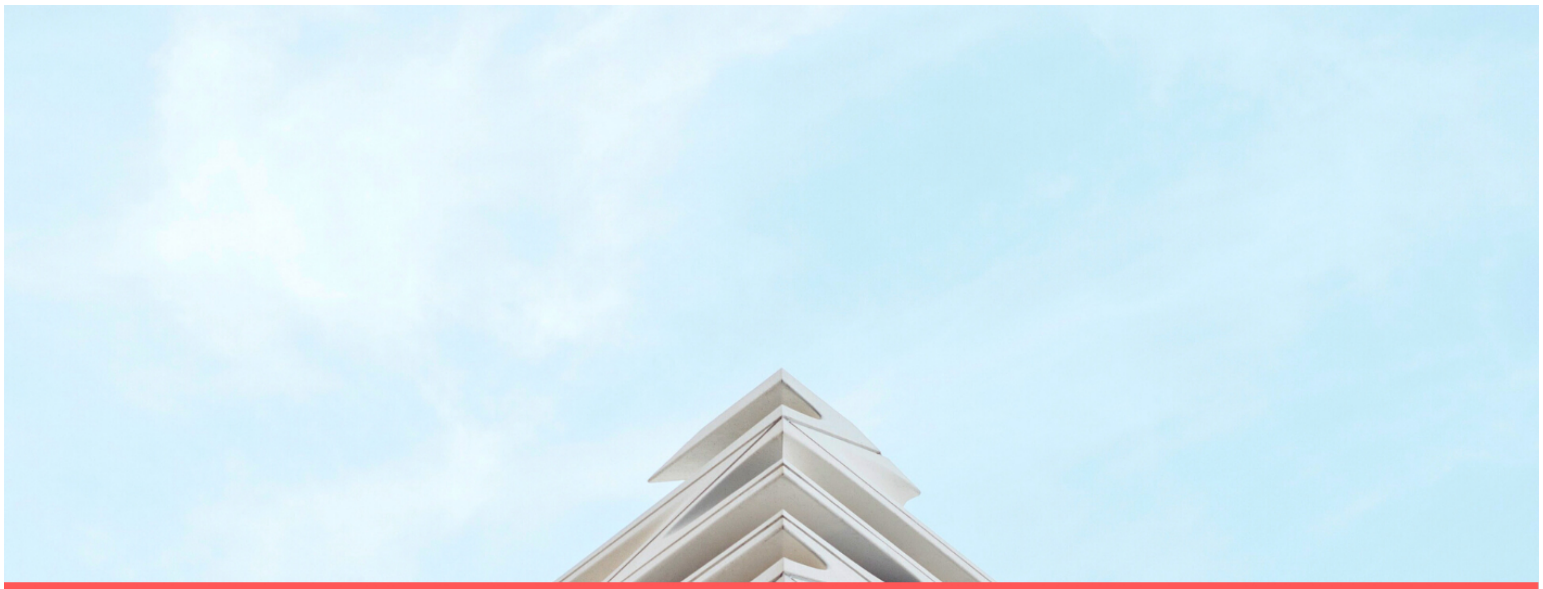


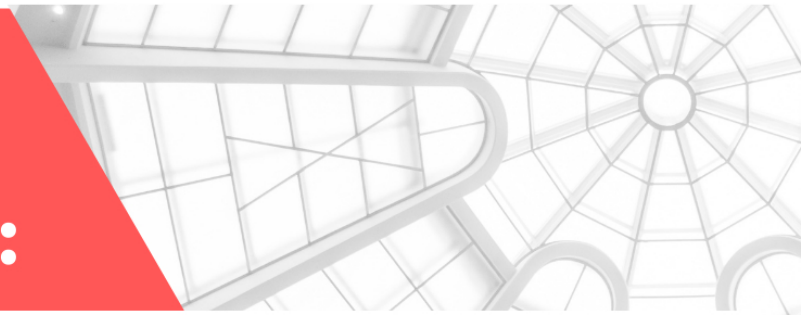
QUORA QUESTIONS MVP

PREPARED BY

Alaa AL-Ghamdi, Arwa Alolyani , Nadia Hajrasi

PRESENTED BY

Submit to: Mr.Ali El-kassas



CASESTUDY:

In this project, we deal with data that contains more than 400,000 rows and 3 columns.
The data is a pair of questions collected from the Quara Questions website.

PROCESSES :

- Knowing and manipulating null values
- Knowing Frequently Asked Questions
- Know the number of frequently asked questions
- Engineering Features:

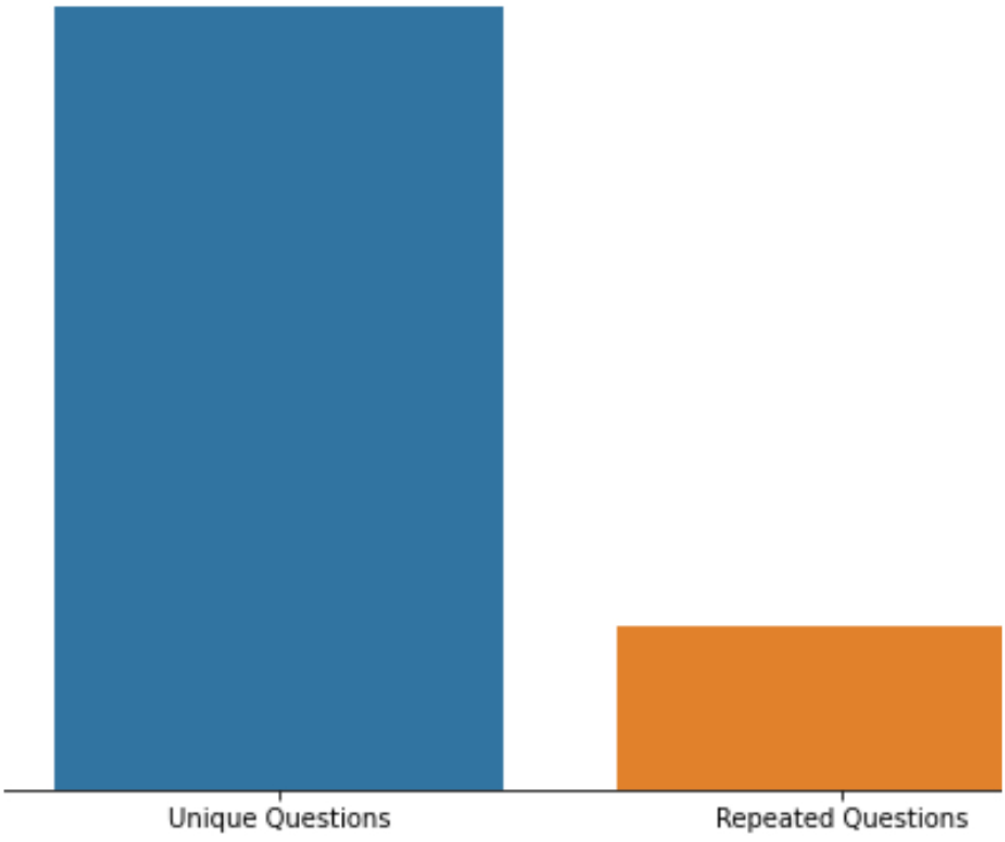
1. Adding new columns containing the number of words in each question
2. Adding new columns showing the length of the question
3. Adding new columns showing the number of words shared in the two questions

- text capture
- Text processing, which includes (punctuation marks, numbers, extreme letters, stop words...)
- Applying machine learning using topic modeling to questions

DATA AFTER ADDING NEW COLUMNS:

question1	question2	is_duplicate	q1len	q2len	q1_num_words	q2_num_words
at are the risks of hormone replacement ther...	Why is hormone replacement therapy a risk?	0	50.0	42.0	8	7
you tell if a person is gay?	How do I know your gay?	1	32.0	23.0	8	6
do I clear IBPS PO in one month?	How do I prepare IBPS PO exam in one month?	1	36.0	43.0	9	10
ere the major effects of the cambodia ea...	What were the major effects of the cambodia ea...	1	124.0	121.0	21	21
t is the funniest joke you've ever heard or...	What is the best joke you have ever heard?	1	52.0	42.0	10	9

FREQUENTLY ASKED QUESTIONS:



QUESTIONS COLUMN 1 :

QUESTIONS COLUMN 2 :

AFTER TEXT PROCESSING (SMALL PATR) :

```
'know gay', 'prepare ibps po exam one month', 'major
ever heard', 'sean combs accumulate wealth', 'best co
xam preparation', 'hospitals first originate', 'like
ht loss program', 'health benefits wasabi paste', 'pr
```

AFTER CONVERTING TEXTS TO NUMBERS :

	abercrombie	actually	add	address	alarm	alcohol	although	amd	america	americac	...	wish	without	word	work	works	world
best smartphone inr	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
prepare abercrombie fitch group interview	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
increase traffic site suggestions get	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
companies inenglanddoes prefer give sponsorship international students	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
best stocks invest	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

TOPIC MODELING TO QUESTIONS :

	Topic0	Topic1	Topic2	Topic3	dominant_topic
Doc0	-0.000000	-0.000000	0.000000	-0.000000	0
Doc1	0.070000	-0.010000	0.010000	0.020000	0
Doc2	0.100000	0.480000	0.140000	0.000000	1
Doc3	0.050000	0.120000	-0.010000	0.250000	3
Doc4	0.020000	0.020000	0.060000	0.050000	2
Doc5	0.000000	0.000000	0.000000	0.000000	0
Doc6	1.020000	0.230000	-0.400000	0.340000	0
Doc7	0.000000	0.000000	0.000000	0.000000	0
Doc8	0.510000	-0.100000	0.160000	-0.160000	0
Doc9	-0.000000	-0.000000	-0.000000	0.000000	0