IS THAT A DUPLICATE QUORA QUESTION?

Team
Sharad Shriyan
Ketan Chaudhari
Sejal Shah
Nikhil Gupta
Ramesh Thota

AGENDA

- Problem
- Train & test data
- Analyzing the data
- Vectorizing the data
- Extra feature selection
- Results

PROBLEM

- Given a pair of questions q1 and q2 we need to determine if they are duplicates of each other.

- More formally:Build a model that learns the function:

  f(q1, q2) = 1 or 0

DUPLICATE QUESTIONS

- How does Quora quickly mark questions as needing improvement?
- Why does Quora mark my questions as needing improvement/clarification before I have time to give it details? Literally within seconds…

- What practical applications might evolve from the discovery of the Higgs Boson?
- What are some practical benefits of discovery of the Higgs Boson?

- Why did Trump win the Presidency?
- How did Donald Trump win the 2016 Presidential Election?

NON-DUPLICATE QUESTIONS

- Who should I address my cover letter to if I'm applying for a big company like Mozilla?
- Which car is better from safety view?""swift or grand i10"".My first priority is safety?
- Mr. Robot (TV series): Is Mr. Robot a good representation of real-life hacking and hacking culture? Is the depiction of hacker societies realistic?
- What mistakes are made when depicting hacking in ""Mr. Robot"" compared to real-life cybersecurity breaches or just a regular use of technologies?
- How can I start an online shopping (e-commerce) website?
- Which web technology is best suitable for building a big E-Commerce website?

THE DATA

This dataset consists of over 400,000 lines of potential question duplicate pairs.

	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

Train data

Question 1 - Question 2 - Answer
Question 3 - Question 4 - Answer

…

Question 400.904 - Question 400.905 - Answer

Test data

Question 1 - Question 2
Question 3 - Question 4

…

Question 2.000.108 - Question 2.000.109

Example

Could time travel ever be possible? - Will time travel ever be possible? - **1**

Why are n't blueberries blue? - Do rubber ducks quack? - **0**

ANALYZING THE DATA

Needed to answer the question: How can a computer determine if two questions are duplicates?

What features makes a pair of questions more likely to be duplicates?

FEATURE ENGINEERING

BASIC FEATURES:

- Length of question 1
- Length of question 2
- Difference between len Q1 and len Q2
- Characters length of question 1
- Characters length of question 2
- Words length of question 1
- Words length of question 2

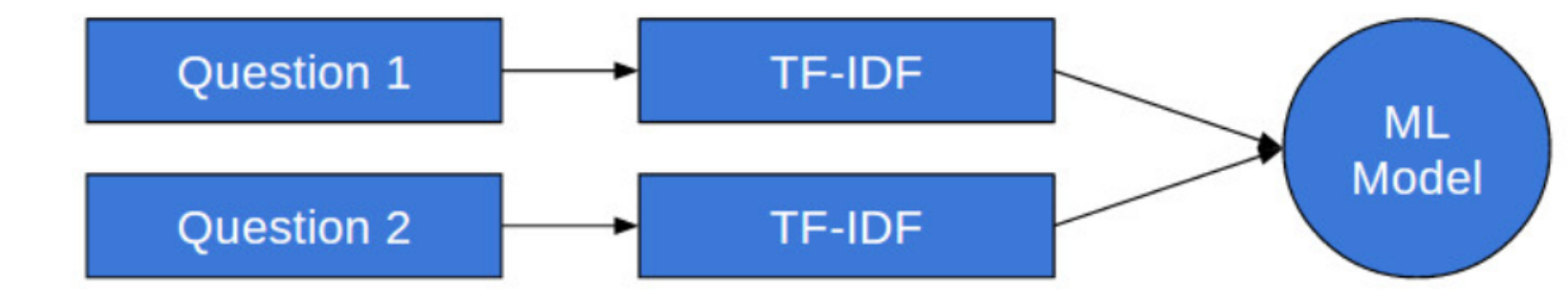
DISTANCE FEATURES:

- Cosine distance

FUZZY FEATURES

- QRatio (This is used)
- WRatio
- Token set ratio
- Token sort ratio

TF-IDF FEATURES WITH PADDING



MODEL-1 - VECTORIZING THE DATA

	simi_score	fuzzy_wuzzy	len_q1	len_q2	diff_len	len_char_q1	len_char_q2	len_word_q1	len_word_q2	common_words
0	0.908893	93	66	57	9	20	20	14	12	10
1	0.798935	65	51	88	-37	21	29	8	13	4
2	0.845154	45	73	59	14	25	24	14	10	4
3	0.809693	7	50	65	-15	19	26	11	9	0
4	0.700268	37	76	39	37	25	18	13	7	2

MODEL - 1 - RESULTS

For 1000 records

- Log loss - 0.595629942077
- Accuracy - 63.6%

For 5000 records

- Log loss - 0.559801468258
- Accuracy - 68.64%

WORD2VEC

FEATURES

- Multi-dimensional vector for all the words in any dictionary
- Always great insights
- Very popular in natural language processing tasks

```
s1_afv = avg_feature_vector('what is step by step guidence for share market?', model=model, num_features=100, index2
s2_afv = avg_feature_vector('what is step by step guidence for share market india?', model=model, num_features=100,
sim = 1 - spatial.distance.cosine(s1_afv, s2_afv)
print(sim)
```

```
[ 1.11654389e-03  2.01525702e-03  2.46455730e-03  1.76973466e-03
  1.80303177e-03  3.13298427e-03  2.67489324e-03  4.76201152e-04
  1.11062953e-03 -2.46574264e-03  6.53219584e-04 -2.11150409e-03
  9.49954614e-04 -3.83553817e-03 -4.35267109e-03  1.71172351e-03
  2.59907776e-03 -2.73996941e-03 -4.10076138e-03 -2.27421639e-03
  3.79554578e-03 -1.44018477e-03 -3.65757197e-03  1.63901981e-03
  8.16443004e-04 -3.05498298e-03 -2.48412765e-03 -1.14748732e-03
  9.47072578e-04 -4.10823710e-03 -4.20046970e-04 -4.67843749e-03
  2.05882126e-03 -2.05447245e-03 -5.80357097e-04 -2.76257959e-03
  3.74685740e-04  1.29257853e-03 -3.83390393e-03 -1.14434304e-04
  1.11643958e-03  5.07403165e-04 -2.82259658e-03  1.19827846e-05
  2.69336160e-03  1.75268622e-04  4.31432854e-03  1.16631715e-03
 -1.02495169e-03 -3.52945295e-03  1.40808325e-03  1.39002816e-03
  2.68074614e-03 -1.76780426e-03  9.99784213e-04  1.72511011e-03
 -1.63186621e-03  5.87369257e-04 -1.97370513e-03  3.23773269e-03
  2.10117386e-03  9.20839608e-04  2.22865120e-03 -3.17564048e-03
  1.19262282e-03  3.27540562e-03 -2.82456982e-03  1.26740942e-03
  2.60676607e-03 -4.60581687e-05  1.84460636e-03  2.36162264e-03
 -3.56886350e-03  2.12654821e-03 -7.63840333e-04  2.19163441e-04
  2.11134017e-03  7.11451808e-04  2.02332414e-03 -2.61263736e-03
 -1.21964887e-03  2.43620109e-03  2.33951001e-03  1.18366920e-03
```

```
 -8.24527116e-04  3.12784407e-03 -1.72695227e-03  3.19600548e-03
 -1.49858149e-03 -2.63706222e-03 -2.05447595e-03  3.48979840e-04
  3.49192019e-03 -9.01978172e-04 -8.27278185e-04 -2.66046287e-03]
[ 1.4824150e-03  2.3410304e-03  2.5057201e-03  7.1959221e-04
  2.5589564e-03  2.0023633e-03  2.6256428e-03 -6.4743502e-04
  1.9883062e-03 -2.3621202e-03  8.1046764e-04 -4.7703227e-04
  1.4472968e-04 -2.7010776e-03 -2.1020719e-03  2.3707305e-03
  1.1218218e-03 -1.4584893e-03 -2.2041420e-03 -6.1709993e-04
  2.6328112e-03 -2.0198794e-03 -3.6344407e-03  1.6273645e-03
  3.2465148e-04 -1.0746701e-03 -2.6007481e-03 -1.3023838e-03
  9.6782576e-04 -2.9667374e-03 -3.3029082e-04 -3.2230695e-03
  8.6386054e-04 -2.5841538e-03 -1.3536937e-04 -2.4909854e-03
 -9.2242815e-04  1.5814705e-03 -3.1863218e-03  1.1359200e-04
  1.6111601e-03  1.2171073e-03 -1.1976762e-03 -1.0747432e-03
  3.0367870e-03  9.8100980e-04  3.6083527e-03  6.3168351e-05
 -1.1800911e-03 -1.6606941e-03  7.3699729e-04  4.7797617e-04
  1.5365584e-03 -2.5491500e-03  1.8692118e-03  1.6889602e-03
 -5.6949531e-04 -2.4491898e-04 -1.0398775e-03  2.7091943e-03
  2.0392933e-03  5.5524462e-04  1.5022462e-03 -3.5042511e-03
  1.8009318e-03  3.6806709e-03 -2.2202660e-03  1.1596442e-03
  8.0664130e-04  3.9891855e-04  1.8341009e-03  2.2164891e-03
 -1.6011344e-03  9.9646987e-04 -1.3051194e-03  1.7954258e-04
  1.2977254e-03  5.1995245e-04  6.2358013e-04 -1.5270450e-03
  2.8931233e-04  1.9982764e-03  2.1500536e-03  1.5250258e-03
  2.8320730e-03 -2.0453138e-03 -1.3015746e-03 -7.2834274e-04
 -1.2974272e-04  3.2787693e-03 -1.3111136e-03  1.6699005e-03
 -6.9379469e-04 -1.0039076e-03 -1.9353554e-03  5.0903851e-04
  3.6861855e-03 -1.2270965e-03 -5.7218940e-04 -1.0998166e-03]
0.9177963137626648
```

Any questions?

THANK YOU