ML в задаче поиска повторяющихся вопросов

О себе

- Учусь на ИМИТ (матфаке) ОмГУ
- Работаю в 7bits
- Преподаю в школе программиста

Your home for data science

train.csv

test.csv

Overview Data Kernels Discussion **Leaderboard** More My Submissions **Submit Predictions**

1	—	✱ DL guys	0.11773	212	8h
2	—	✱ Depp Learning	0.11869	143	21h
3	—	✱ Jared Turkewitz & sjv	0.11931	123	10h
4	▲ 2	YesOfCourse	0.12263	40	17h
5	▼ 1	No Questions Asked	0.12358	201	3h
6	▼ 1	I can't speak English.	0.12382	171	20m
7	—	Qingchen & Co	0.12406	161	20h
8	—	raddar & clustifier & Abhishek	0.12502	245	2h
9	—	NLPFakers	0.12518	212	3h
10	▲ 3	AMAA power	0.12622	209	20h
11	▼ 1	aphex34	0.12687	124	4h
12	▼ 1	Quora Gold	0.12702	229	3h
13	▲ 18	Why so duplicated?	0.12716	201	2h
14	▼ 2	HouJP	0.12831	78	4h

Постановка задачи

train.csv

qid1	qid2	question1	question2	is_duplicate
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
3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0
11	12	Astrology: I am a Capricorn Sun Cap moon and c...	I'm a triple Capricorn (Sun, Moon and ascendan...	1
13	14	Should I buy tiago?	What keeps childern active and far from phone ...	0
15	16	How can I be a good geologist?	What should I do to be a great geologist?	1

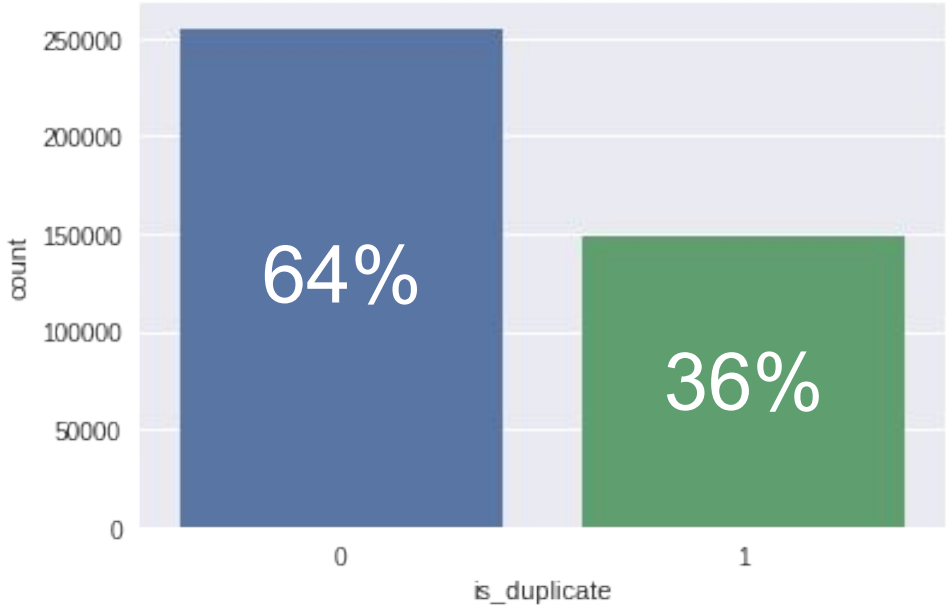
test.csv

	test_id	question1	question2
0	0	How does the Surface Pro himself 4 compare wit...	Why did Microsoft choose core m3 and not core ...
1	1	Should I have a hair transplant at age 24? How...	How much cost does hair transplant require?
2	2	What but is the best way to send money from Ch...	What you send money to China?
3	3	Which food not emulsifiers?	What foods fibre?
4	4	How "aberystwyth" start reading?	How their can I start reading?
5	5	How are the two wheeler insurance from Bharti ...	I admire I am considering of buying insurance ...
6	6	How can I reduce my belly fat through a diet?	How can I reduce my lower belly fat in one month?
7	7	By scrapping the 500 and 1000 rupee notes, how...	How will the recent move to declare 500 and 10...

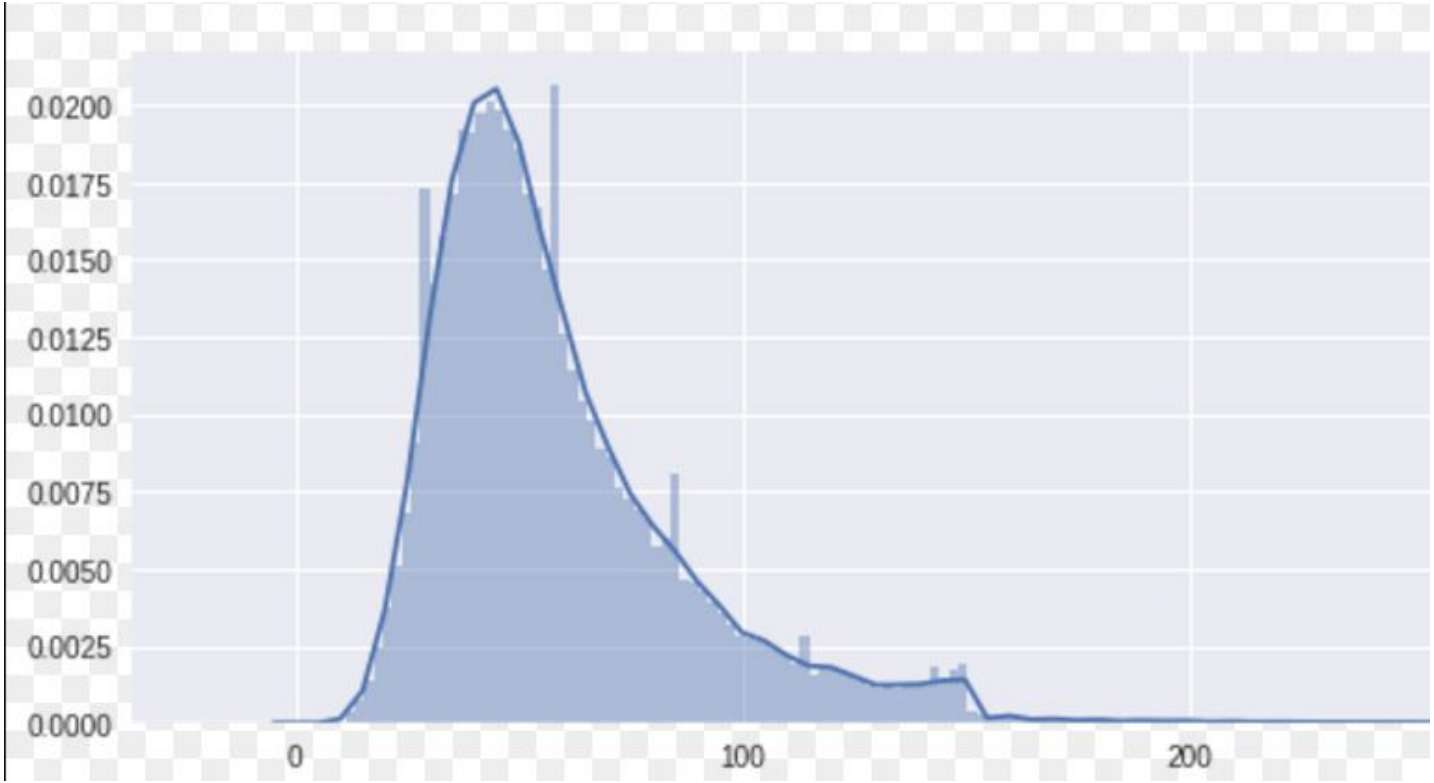
Score: log loss

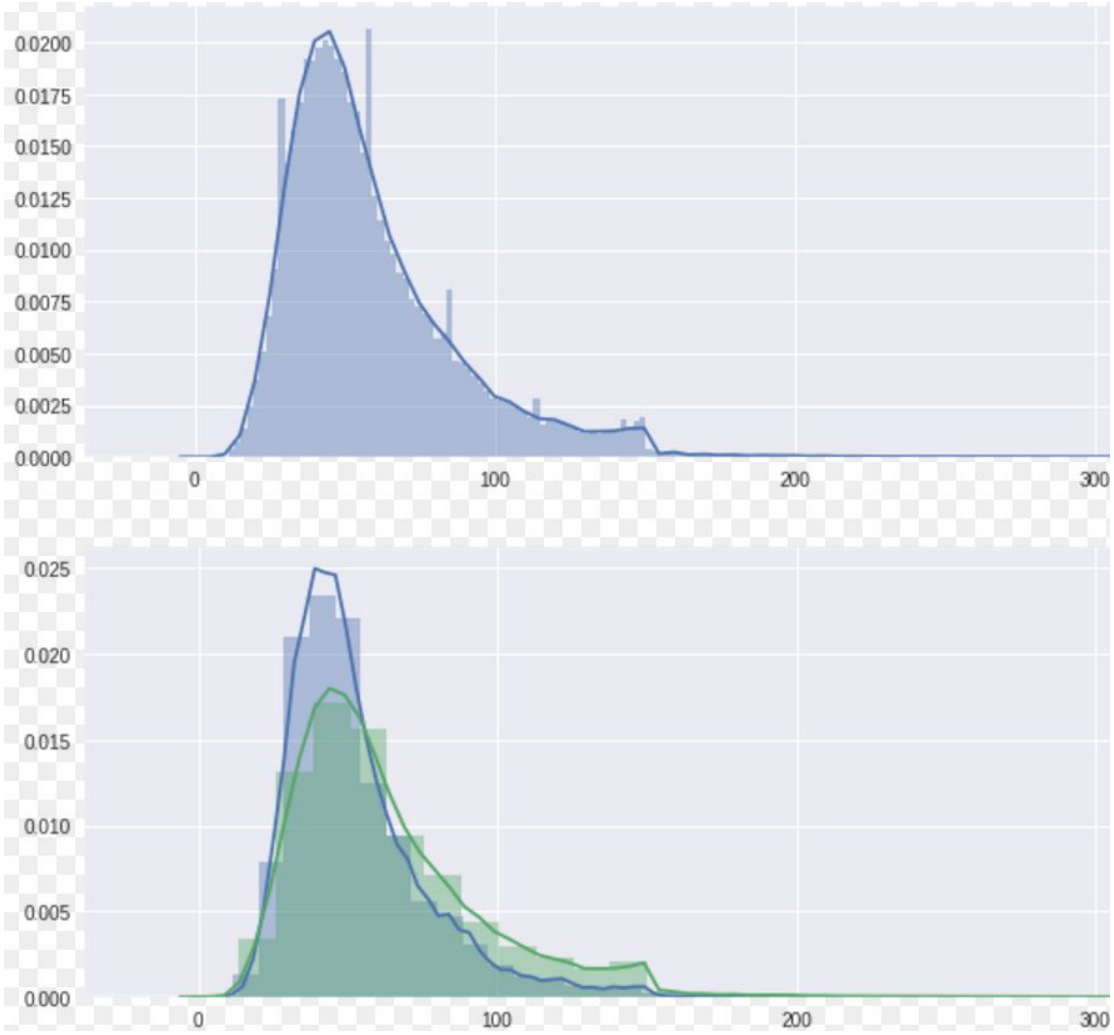
$$-\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)].$$

Количество повторяющихся вопросов



Длина вопросов



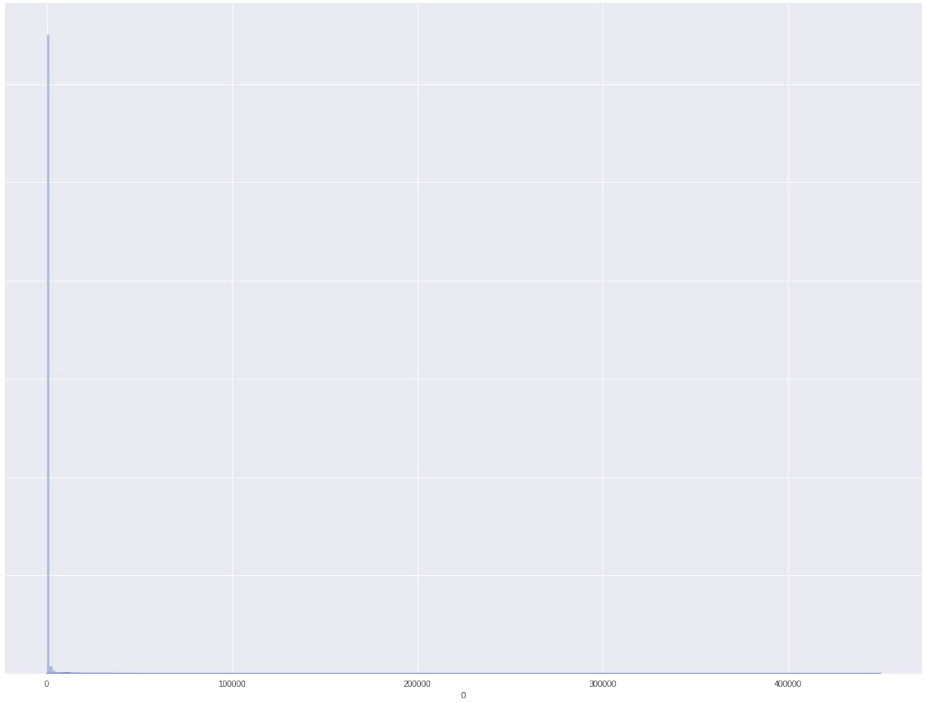

0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
100
200
300
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300

Почему не bag-of-words?

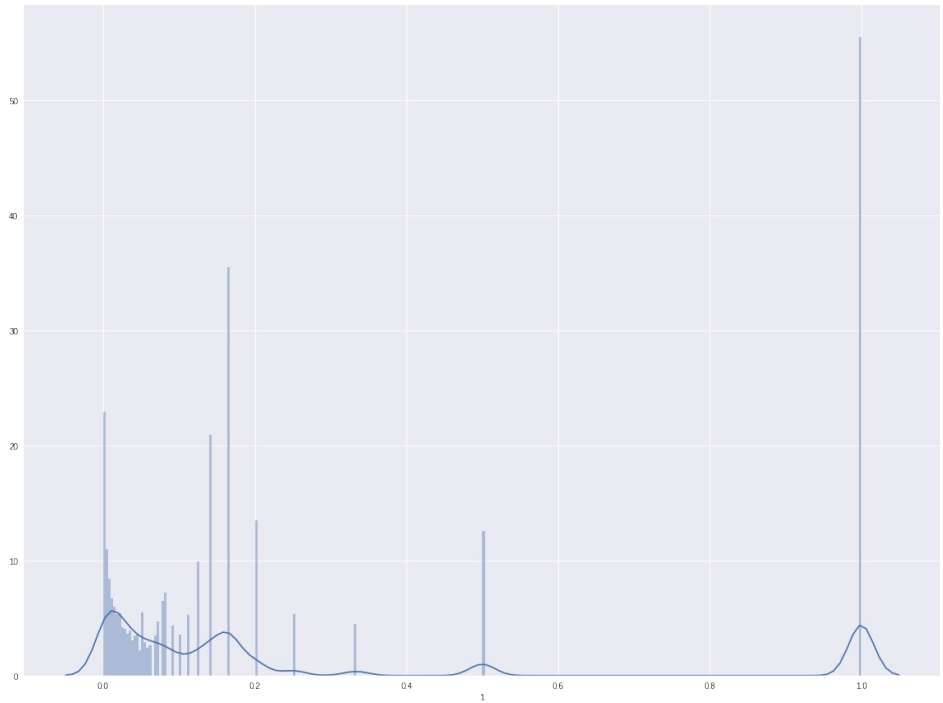
Предложение - множество слов

$Q1 \cap Q2$ $Q1 \cup Q2$ $Q1 \oplus Q2$

Взвешивание слов



count(word)



log(count(word))

TF-IDF

TF - term frequency

IDF - inverse document frequency

$$\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k}$$

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|}$$

$$\mathrm{tf\text{-}idf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

Слова - векторы

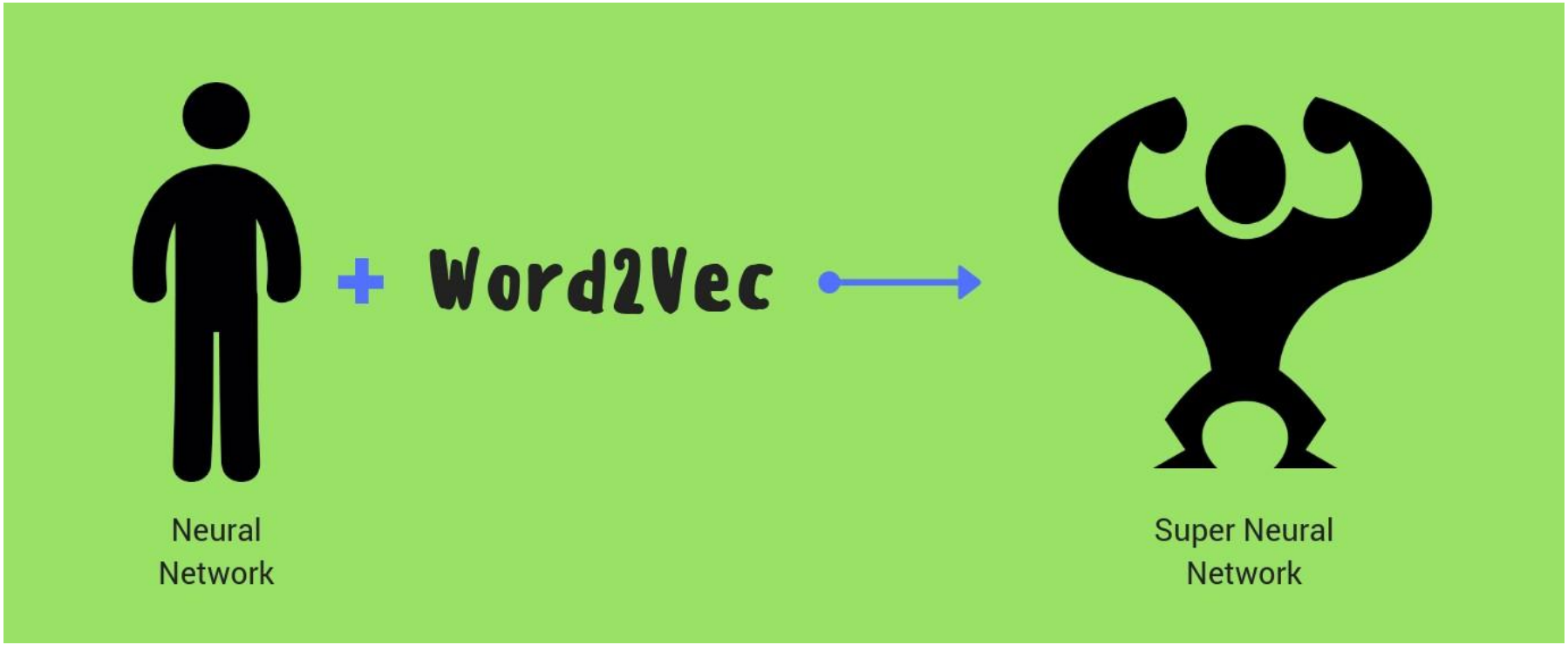
'How can I be a good geologist?'

'What should I do to be a great geologist?'

'How can I see all my Youtube comments?'

	all	be	can	comments	do	geologist	good	great	how	my	see	should	to	what	youtube
0	0	1	1	0	0	1	1	0	1	0	0	0	0	0	0
1	0	1	0	0	1	1	0	1	0	0	0	1	1	1	0
2	1	0	1	1	0	0	0	0	1	1	1	0	0	0	1

Word2Vec



Неудачные признаки

Расстояние Джаккарда $$\rho_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Неудачные признаки

Расстояние Джаккарда $\rho_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$

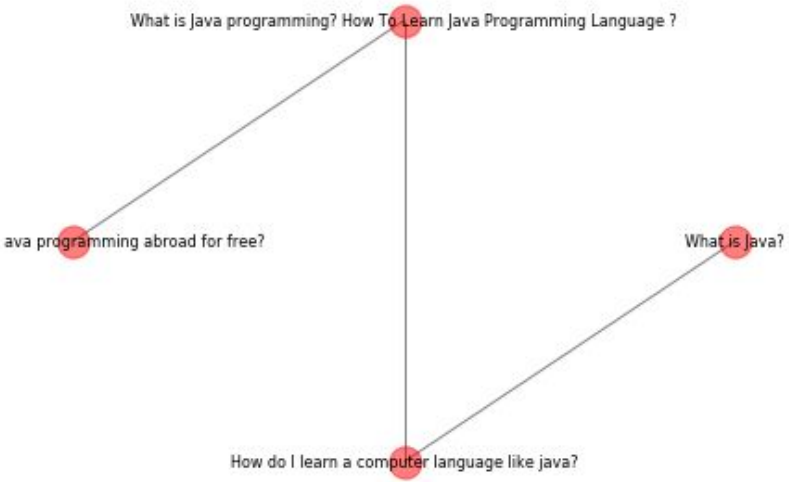
F(Q1, Q2) = α1*(Q1 ∩ Q2) + α2*(Q1 U Q2) + α3*(Q1⊕Q2)

То, что не успел:

N-gram

Doc2vec

Кластеризация



Особенности Kaggle

Разное распределение “положительных” пар в train и test (36% и 17%)

$$f(x) = \frac{\gamma_1 x}{\gamma_1 x + \gamma_0 (1 - x)}$$

γ0 = 1.30905513329
γ1 = 0.472008228977

Особенности kaggle

Влияние ID

train.csv

qid1	qid2	question1	question2	is_duplicate
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
3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0
11	12	Astrology: I am a Capricorn Sun Cap moon and c...	I'm a triple Capricorn (Sun, Moon and ascendan...	1
13	14	Should I buy tiago?	What keeps childern active and far from phone ...	0
15	16	How can I be a good geologist?	What should I do to be a great geologist?	1

test.csv

	test_id	question1	question2
0	0	How does the Surface Pro himself 4 compare wit...	Why did Microsoft choose core m3 and not core ...
1	1	Should I have a hair transplant at age 24? How...	How much cost does hair transplant require?
2	2	What but is the best way to send money from Ch...	What you send money to China?
3	3	Which food not emulsifiers?	What foods fibre?
4	4	How "aberystwyth" start reading?	How their can I start reading?
5	5	How are the two wheeler insurance from Bharti ...	I admire I am considering of buying insurance ...
6	6	How can I reduce my belly fat through a diet?	How can I reduce my lower belly fat in one month?
7	7	By scrapping the 500 and 1000 rupee notes, how...	How will the recent move to declare 500 and 10...

Вопросы?