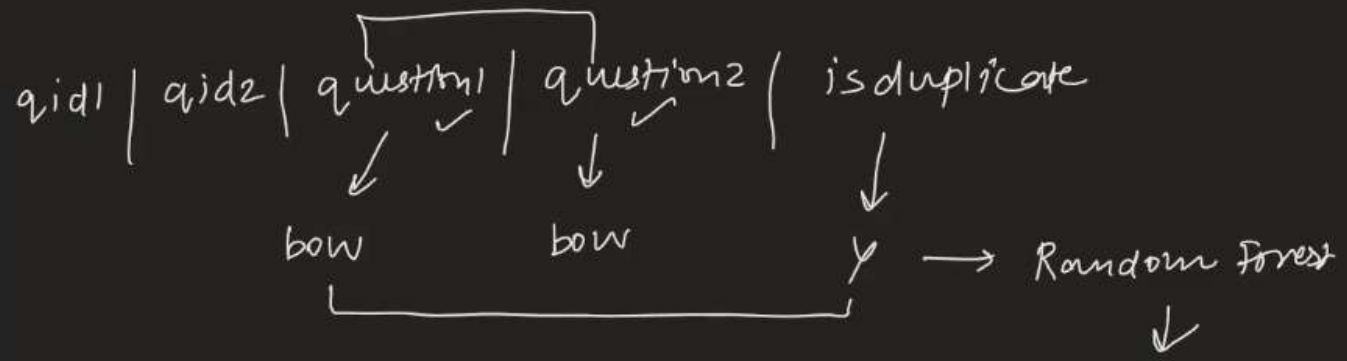
Friday, February 11, 2022 7:05 AM



1. q1 len → char length of q1
2. q2 len → char length of q2
3. q1 words → # words in q1
4. q2 words → # words in q2
5. words common → # of common unique words
6. words total → Total # words in q1 + Total # of words in q2
7. word share → word common / word total

tokens, words, stopwords is of
5 3 2
[sachin] [is] [a] [great] [batsman]

Token Features

1) cwc_min —
2) cwc_max —
3) csc_min —
4) csc_max —
5) ctc_min —
6) ctc_max —
7) last_word_eq
8) first_word_eq

→ # common words / min(words(q1, q2)) — ③ ; ↳ max

[q1 q2] → 3, ④

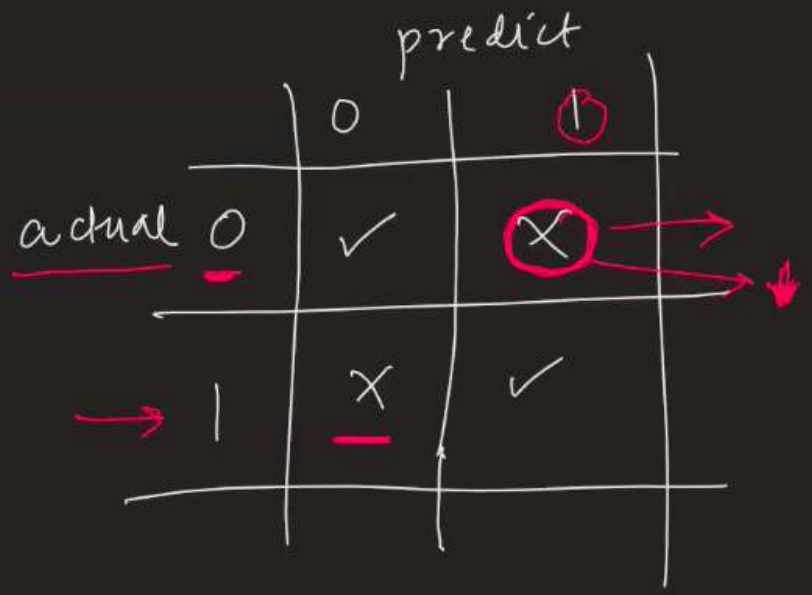
q1	q2
5	6

→ # common stop words = / min(stopwords(q1, q2)) ← 5 ; ↳ max

q1	q2
10	15

→ # common tokens / min(tokens(q1, q2)) ← 10 ; ↳ max

Length Based features



AD → 0
PD → 1

random
xgboost

1) Increase data (42) → 30K → RAM
   - → increase RAM
   - → Cloud platform
   - → incremental learning → Vaex / Dask

2) Preprocessing → Stemming

3) Apply more algorithms
   - → SVM Logistic
   - → Hyperparameter tuning
   - → cross validation

4) More feature (22) →

5) Bag of word
   - → tfidf
   - → word2vec
   - → tfidf weighted W2V

6) Deep learning → 42 → data loaders