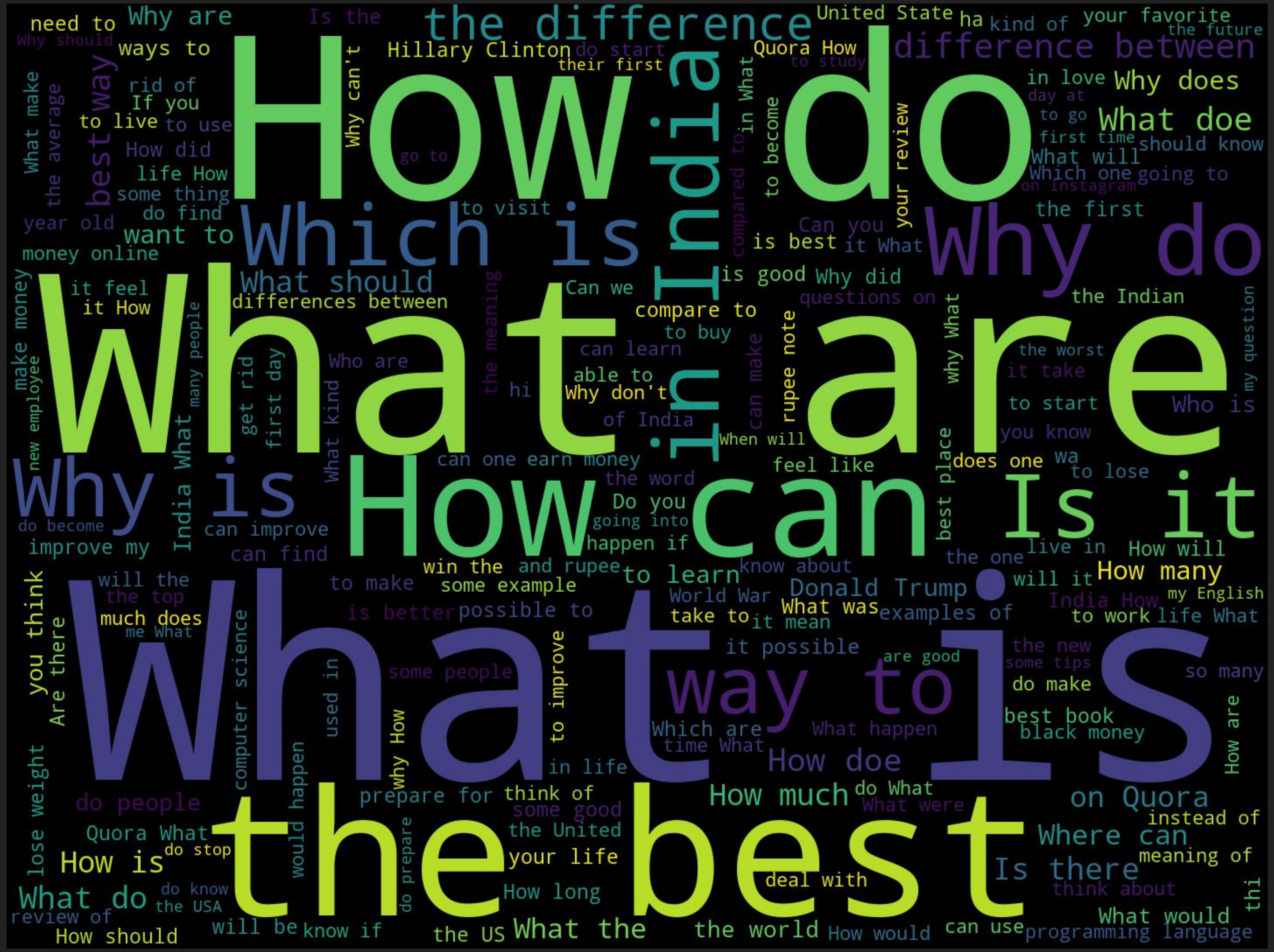
ZHAO JIN

KANGJING FAN

QUOTA QUESTION PAIRS

OVERVIEW

- Task: Identify duplicate questions.
- Dataset size:
  - Training:
    - Size (404290), Field (question pairs+gold label)
  - Test:
    - Size (2345796), Field (question pairs)
  - Character: Range:(15,150), Mean(~60)
  - Word: Range:(3,30), Mean(~11)



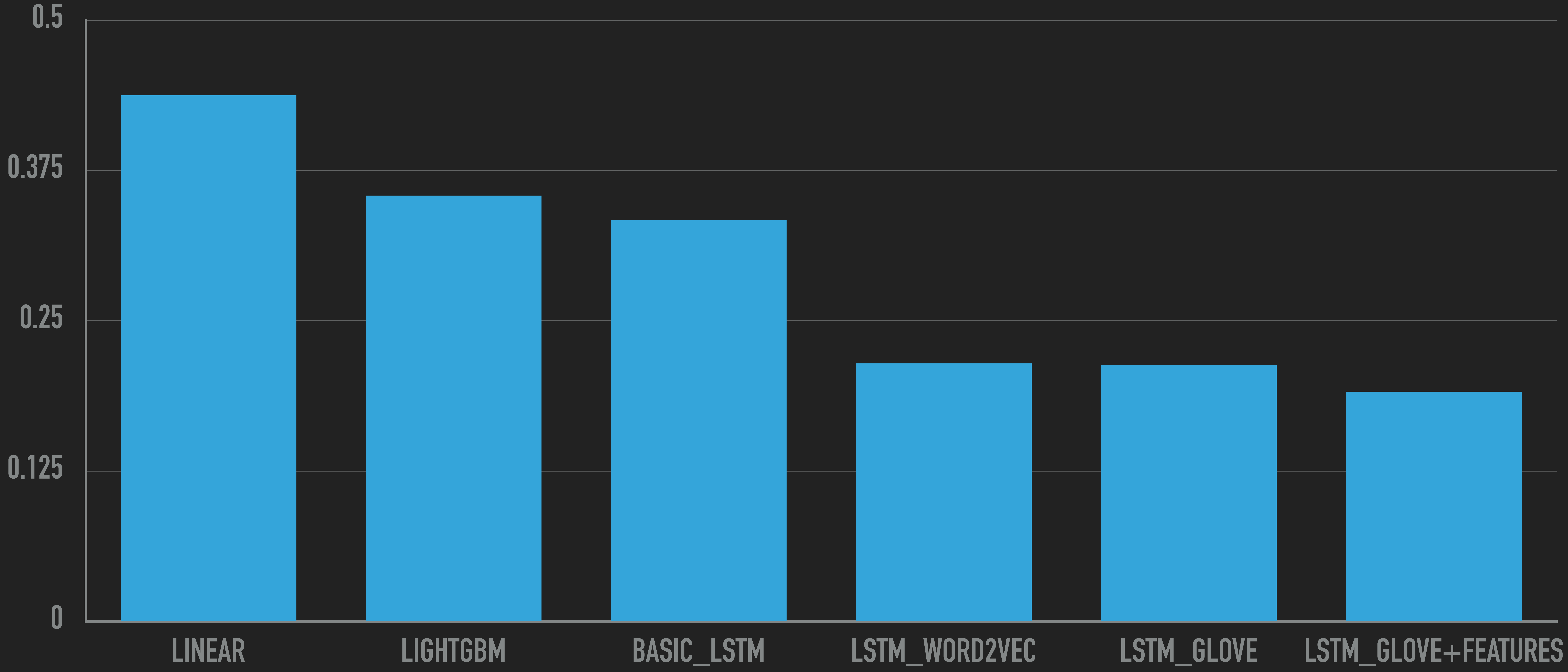
PREPROCESSING

- Use re package separating punctuation from the word, restore abbreviation
  - ’s —- is
- Remove stop-words
- Correct spelling: Spell checker & word2vec
- Lemmatize word (WordNetLemmatizer())
- Get word stem (SnowballStemmer(‘english'))

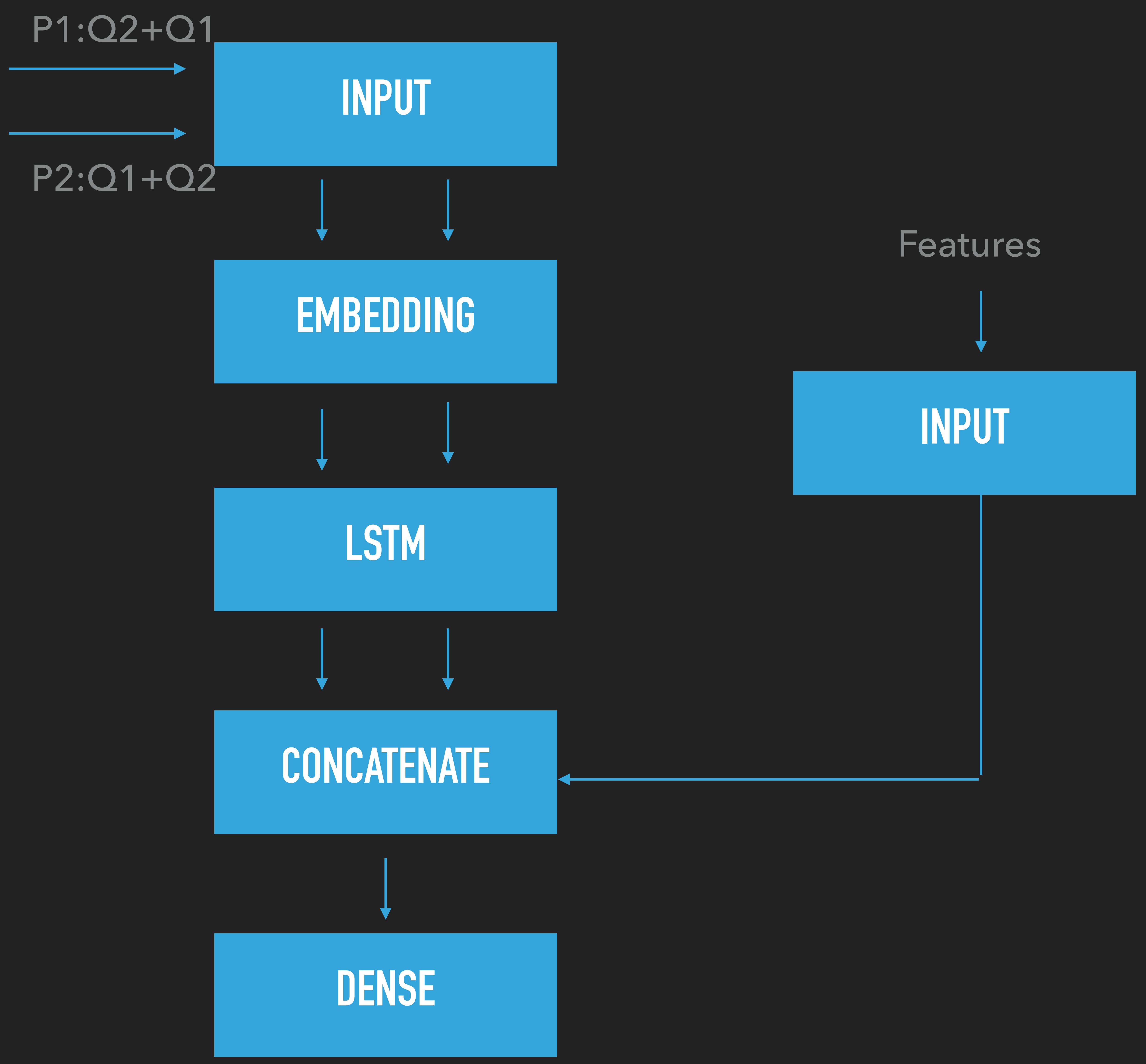
METHOD

- Common word + linear model $\text{len}(Q1 \cap Q2) / \text{len}(Q1 \cup Q2)$
- Magic features + LIGHTGBM
- Basic LSTM
- Boost LSTM (with pre-trained word embedding)
- Boost LSTM + magic features

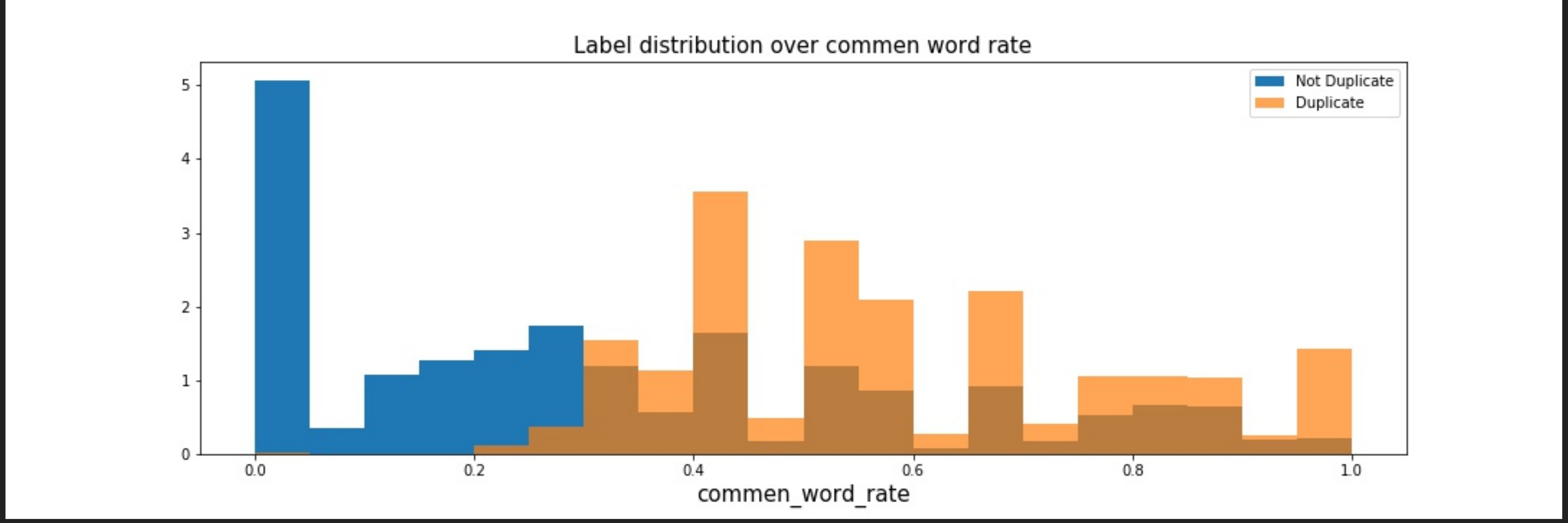
RESULT

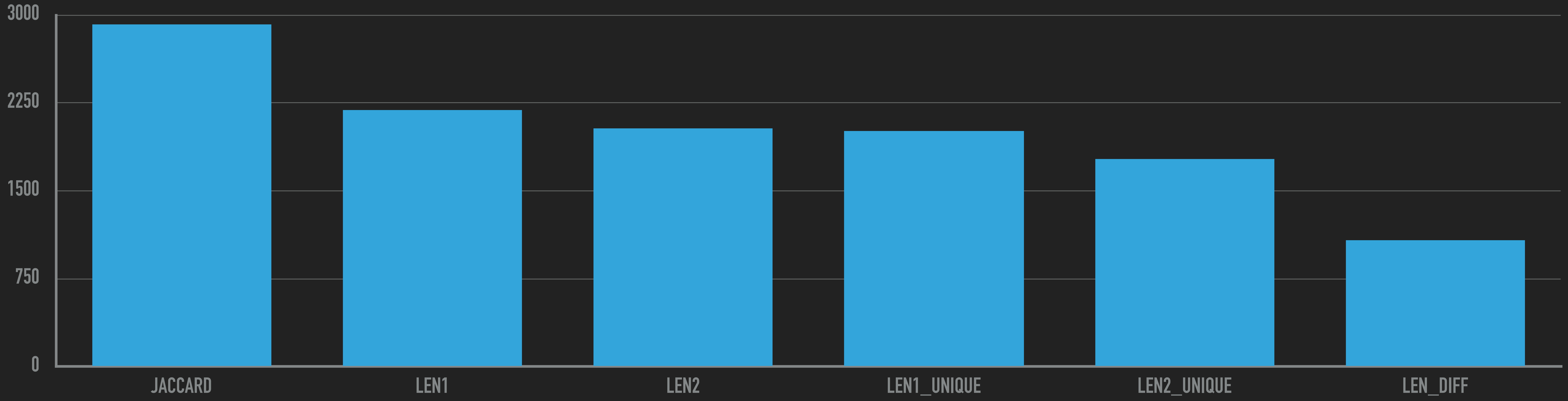

0.5
0.375
0.25
0.125
0
LINEAR
LIGHTGBM
BASIC_LSTM
LSTM_WORD2VEC
LSTM_GLOVE
LSTM_GLOVE+FEATURES

BEST MODEL


P1:Q2+Q1
P2:Q1+Q2
INPUT
EMBEDDING
LSTM
CONCATENATE
DENSE
Features
INPUT

ANALYSIS


Label distribution over commen word rate
Not Duplicate
Duplicate
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
commen_word_rate


3000
2250
1500
750
0
JACCARD
LEN1
LEN2
LEN1_UNIQUE
LEN2_UNIQUE
LEN_DIFF

CONCLUSION

- For the data pre-processing
  - Spell check is necessary
  - Spell check + lemmatize + stem
- For the model
  - LSTM > Tree base model
  - Ensemble model is the best choice

THANK YOU

FEATURES

- Jacquard: $\mathrm{len}(Q1 \cap Q2) / \mathrm{len}(Q1 \cup Q2)$
- Len1: $\mathrm{len}(Q1)$
- Len2: $\mathrm{len}(Q2)$
- Len1_unique: $\mathrm{len}(Q1 - Q1 \cap Q2)$
- Len2_unique: $\mathrm{len}(Q2 - Q1 \cap Q2)$
- Len_diff: $|\mathrm{len}(Q1) - \mathrm{len}(Q2)|$