Quora Question Pairs - Kaggle

•••

Bhavika Tekwani

Problem Description

- 100 million people visit Quora every month.
- People often post the same questions but framed differently.
- We want to make content discovery easier - search & relevance, recommendations, autocomplete, etc
- The goal is to discover whether two questions have the same intent!

Duplicate question detection in the wild

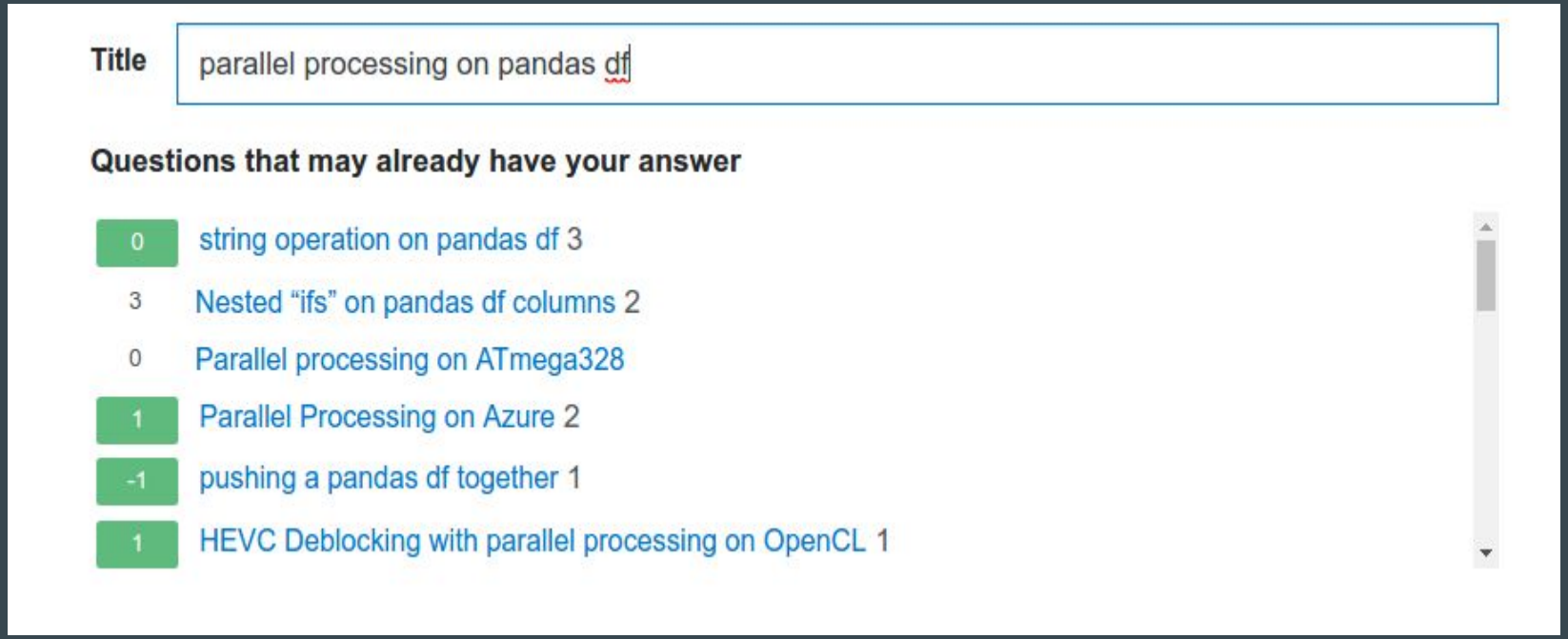


Fig. 1 Stack Overflow - Ask page (05/02/2017)

Duplicate question detection in the wild

Google

parallel processing on pandas df

All News Images Videos Shopping More Settings Tools

About 242,000 results (0.50 seconds)

python - Large Pandas Dataframe parallel processing - Stack Overflow
stackoverflow.com/questions/33612935/large-pandas-dataframe-parallel-processing
Nov 9, 2015 - The entire **DataFrame** needs to be pickled and unpickled for each process created by joblib. In practice, this is very slow and also requires ...

parallel processing in pandas python - Stack Overflow
stackoverflow.com/questions/36054321/parallel-processing-in-pandas-python
Mar 17, 2016 - Here's a webpage I found that might help: ... parallel processing in pandas python · Ask Question ... I have 5,000,000 rows in my dataframe.

Is there a straightforward way to run pandas.DataFrame.isin in parallel?
datascience.stackexchange.com/.../is-there-a-straightforward-way-to-run-pandas-dataf...
May 19, 2014 - Though I know I could manually break up the **dataframe** into chunks and run the operation in ... For **parallelism**, you may be interested in Joblib.

Fig. 2 Google Search results (05/02/2017)

Data

- 404289 question pairs in the train set (64 MB)
- 2345795 question pairs in the test set (314 MB)

```
   id  qid1  qid2                                          question1  \
0   0     1     2  What is the step by step guide to invest in sh...
1   1     3     4  What is the story of Kohinoor (Koh-i-Noor) Dia...
2   2     5     6  How can I increase the speed of my internet co...
3   3     7     8  Why am I mentally very lonely? How can I solve...
4   4     9    10  Which one dissolve in water quikly sugar, salt...

                                           question2  is_duplicate
0  What is the step by step guide to invest in sh...             0
1  What would happen if the Indian government sto...             0
2  How can Internet speed be increased by hacking...             0
3  Find the remainder when [math]23^{24}[/math] i...             0
4            Which fish would survive in salt water?             0
```

Data

- Imbalance: Only 37% of the question pairs are marked as duplicates
- The goal is to predict *is_duplicate* as a value between 0 and 1 for the test set.
- Observation: Some labels in the train set look incorrect.

Evaluation

- Log loss between the predicted values and ground truth. Averaged over N question pairs to generate a leaderboard score.
- The test set has computer generated question pairs as an anti-cheating measure.
- Leaderboard scores are calculated on 35% of the test set.

Feature Engineering

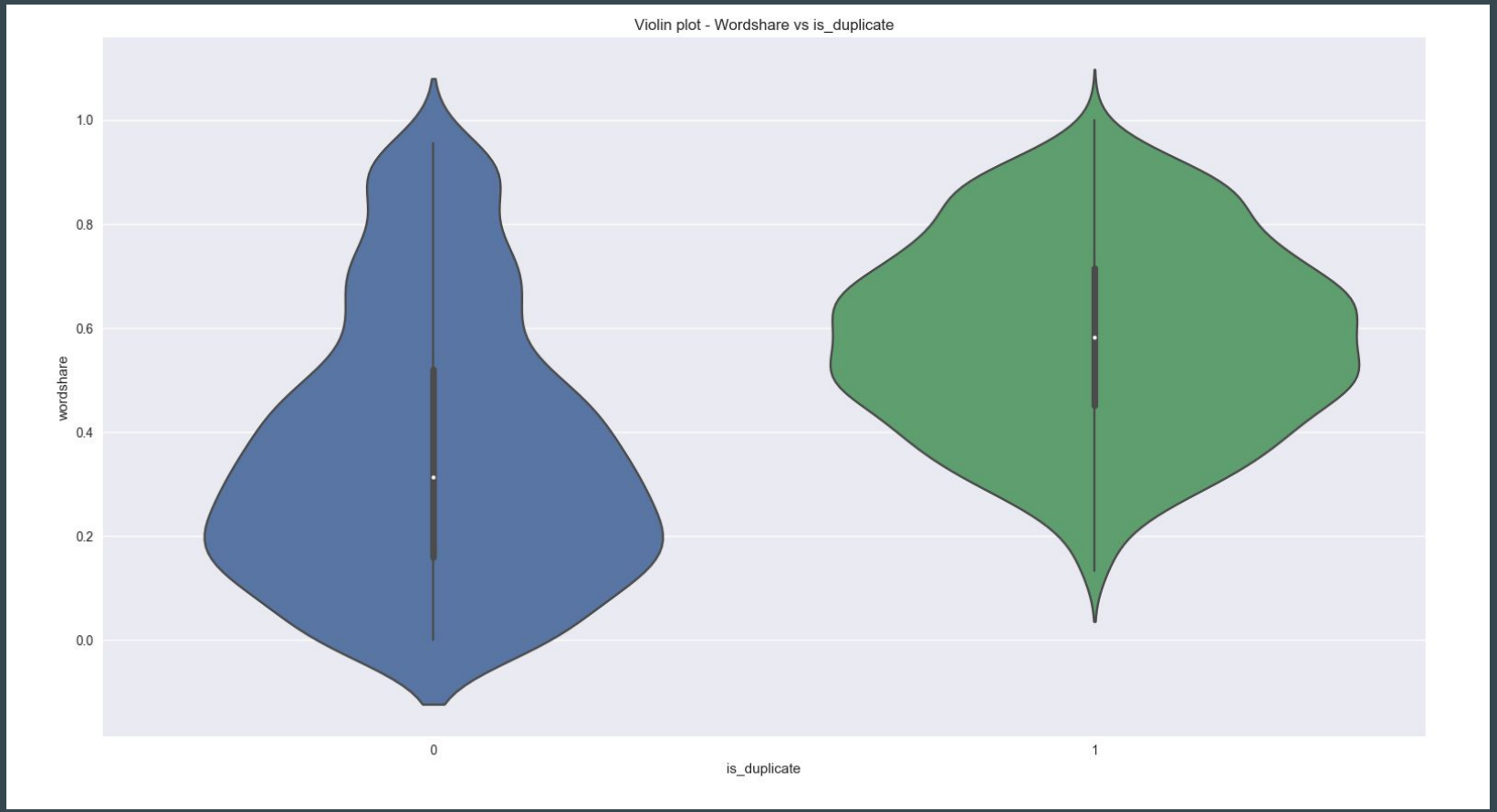
- Preprocessing
  - Stopword removal
  - Corrected typos
  - Tidied up text[1] (“kms” -> kilometres, “the US”, “USA” -> America)
  - Stemming using Porter stemmer
- Features based on lengths and word counts
- Fuzzy features - ratio, partial ratio, token set ratio, token sort ratio, etc
- Wordshare - number of common words in a pair, normalized

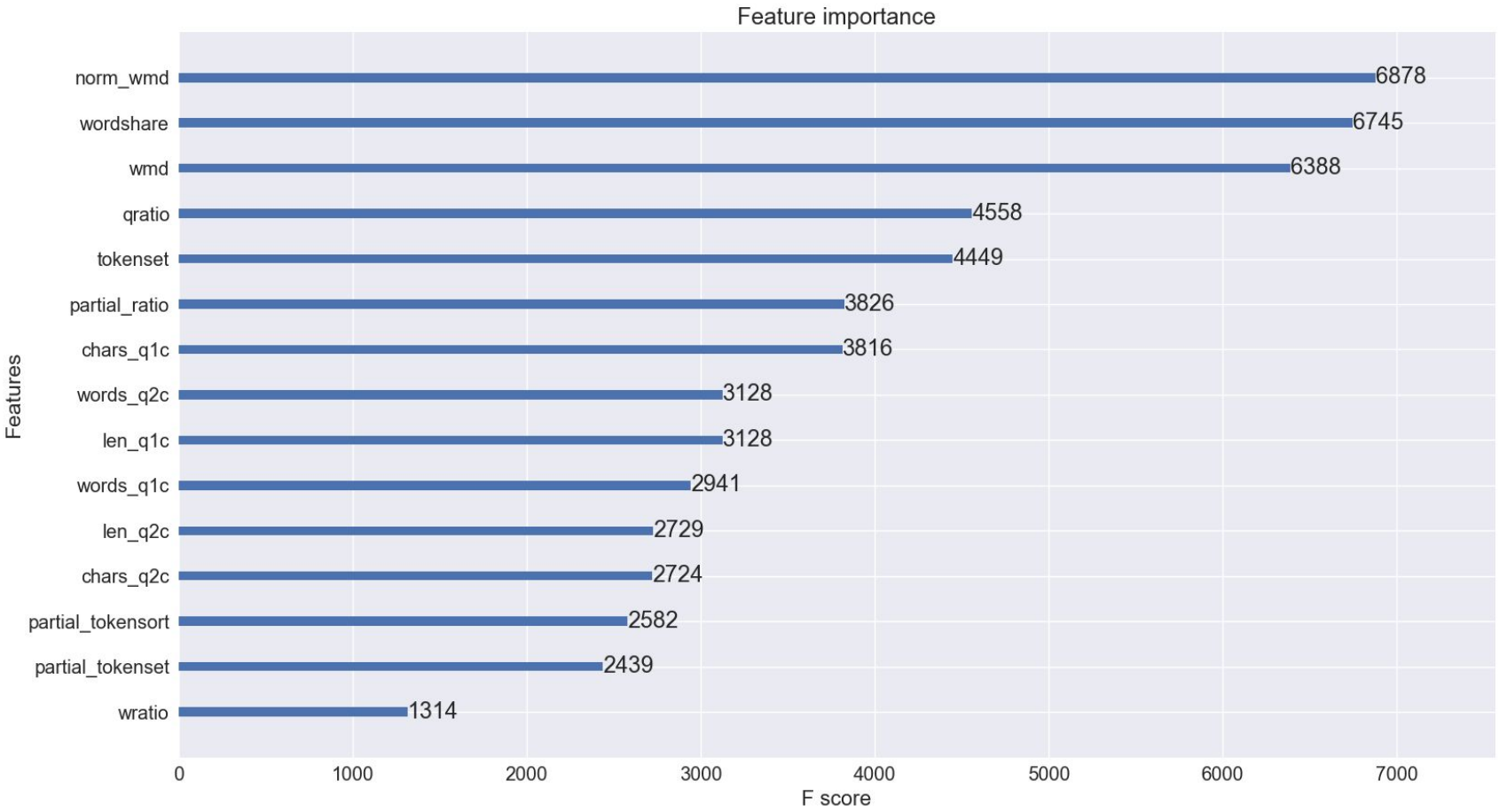
[1] List of corrections created by user ‘currie32’ on Kaggle

Feature Engineering

- Word2Vec
  - Convert each question (in a pair) to a vector of 300 elements - Word2Vec -> Sent2Vec
  - Use these vectors for distance calculations
  - Calculate skew and kurtosis of each vector
- Distance metrics
  - Cosine
  - Euclidean
  - Jaccard
  - Cityblock, etc
- Word Mover's Distance

Which features should we use?




Feature importance
norm_wmd
wordshare
wmd
qratio
tokenset
partial_ratio
chars_q1c
words_q2c
len_q1c
words_q1c
len_q2c
chars_q2c
partial_tokensort
partial_tokenset
wratio
6878
6745
6388
4558
4449
3826
3816
3128
3128
2941
2729
2724
2582
2439
1314
Features
0
1000
2000
3000
4000
5000
6000
7000
F score

Models

Baseline model: XGBoost

Cross validation: 5 fold CV with 80-20 train/test split.

Feature sets:

- All length and word count features - 7 [FS-1]
- Fuzzy features - 6 [FS-2]
- Word Mover's Distance & Normalized Word Mover's Distance - 2 [FS-3]
- Distance metrics, skew and kurtosis - 11 [FS-4]

Results

Model	Features	Log loss (Kaggle)
XGBoost	FS-1 + FS-2 + Wordshare	0.63380
XGBoost [optimized]	Wordshare	0.46868
XGBoost [optimized]	Wordshare + FS-3	0.41594 [#1348]
Blend model - 5 regressors (Random Forest, Extra Trees, Gradient Boosting, Logistic Regression)	FS-1 + FS-2 + FS-3 + Wordshare	0.41859 [#1383]

Further Work

- Train GRUs and LSTMs - compare the performance with XGBoost
- Deal with imbalance in the train set
- Try new features
  - Punctuation - count question marks?
  - POS tags to identify (verb, noun) combinations that are often used in the same context
  - Process text again without removing punctuation

Tools

Visualization: matplotlib, seaborn

Data analysis: numpy, pandas, scikit-learn, XGBoost

NLP: fuzzywuzzy, pyemd, nltk, gensim

Pretrained models: Word2Vec