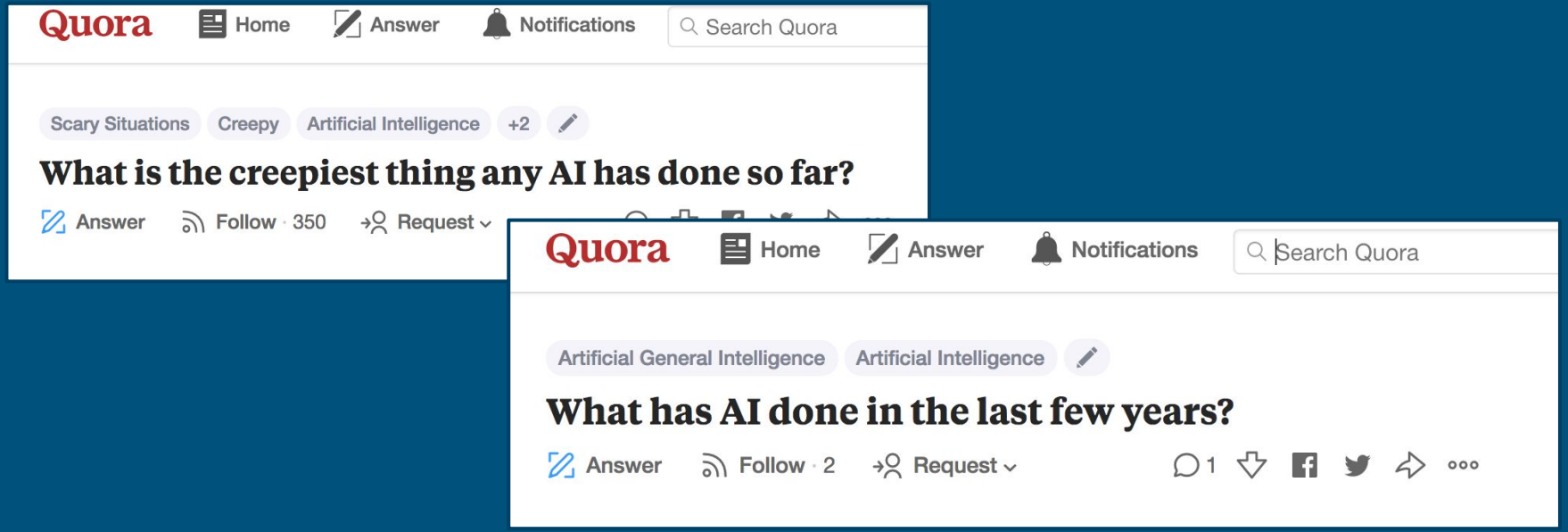
Detecting Duplicate Quora Questions

Emy Parparita

Objective

Predict whether the questions in a pair are a duplicate of each other or not.



Applicability

Forum Sites:

- Better user experience
  - inquirers may get their answers right away
  - respondents do not get annoyed by repeat questions
- Better use of the infrastructure for the provider, may result in cost savings

Input Data

A CSV file provided by Kaggle with the following structure:

```
"id","qid1","qid2","question1","question2","is_duplicate"
"0","1","2","What is the step by step guide to invest in share market in india?","What is the step by step guide to invest in share market?","0"
"20","41","42","Why do rockets look white?","Why are rockets and boosters painted white?","1"
```

and with a caveat:

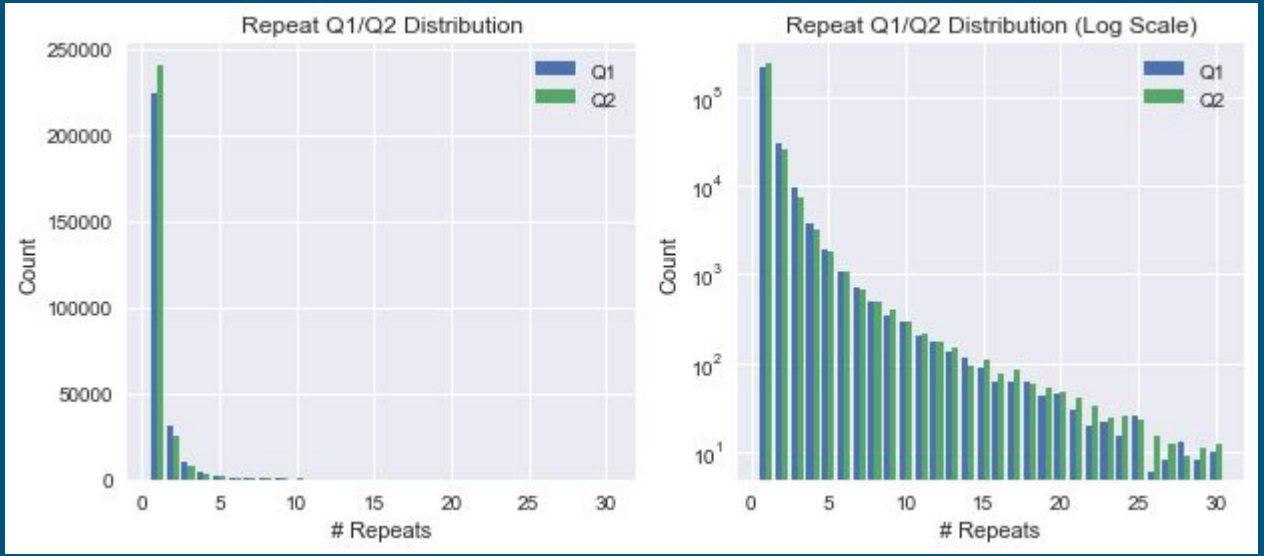
> “The ground truth is the set of labels that have been supplied by human experts. The ground truth labels are inherently subjective, as the true meaning of sentences can never be known with certainty. Human labeling is also a 'noisy' process, and reasonable people will disagree. As a result, the ground truth labels on this dataset should be taken to be 'informed' but not 100% accurate, and may include incorrect labeling.”

Workflow

- Data preparation (one-off):
  - Keep only questions in English
  - Remove empty/duplicate pairs
- Tokenize, stem and filter non-words
- Split 80/20 train/test sets
- Build question -> vector map (LSI, LDA)
- Logistic Regression estimator for vec(q1), vec(q2) -> is_duplicate predictions
- Build similarity vec(q1), vec(q2) -> sim12 map
- Logistic Regression estimator for sim12 -> is_duplicate predictions
- Compute scores: accuracy and log-loss
- Select the best method

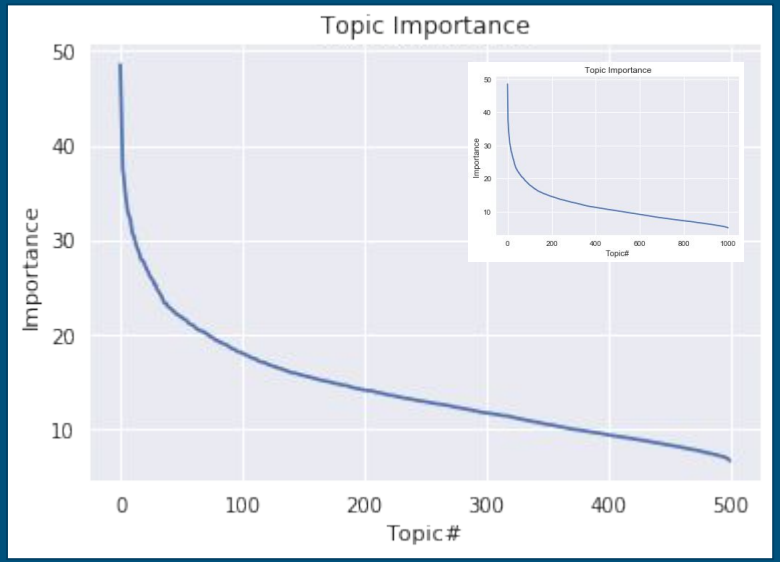
EDA

- 381310 questions
- is_duplicate ratio: .37
- repeat Q1 or Q2 distribution



Number Of Topics

- LSI
  - Start w/ 500 topics and check the elbow plot (see inset for 1000)
  - 200 topics should be enough, however 500 may be tried too if there are enough computing resources
- LDA
  - 200 topics, considering that LDA is better at generating topics



Results And Method Selection

NLP	Num Topics	Estimator	Accuracy (higher is better)	Log Loss (lower is better)
LSI	200	LR on Topics	0.721	0.559
LSI	200	LR on Similarity	0.639	0.625
LSI	500	LR on Topics	0.721	0.558
LSI	500	LR on Similarity	0.643	0.620
LDA	200	LR on Topics	0.717	0.609
LDA	200	LR on Similarity	0.636	0.632
Random guessing 1 with p = .37 (is_duplicate ratio)			0.534	0.659
Kaggle top score (140 times better)				0.11277

What Went Wrong

No synonym awareness:

```
Q1: 'How much equity should I give to CTO?'
Q2: 'How much equity should I offer a CTO?'
is_dup=1, pred=0, sim=0.28923148979392366
```

No meaning awareness:

```
Q1: 'What should a person do when everything goes wrong in their life?'
Q2: 'What should do when nothing goes right in life?'
is_dup=1, pred=0, sim=0.33404120300891316
```

Number exclusion:

```
Q1: 'What are the best Doctor Who episodes with the 10th Doctor?'
Q2: 'What are the best Doctor Who episodes with the 5th Doctor?'
is_dup=0, pred=1, sim=0.998440850819464
```

Future Ideas

- Use Word2Vec because it is better at capturing the context for words
- Cluster Word2Vec representations into synonym groups
- Convert docs into word -> synonym-group representation lists
- Use the converted docs as input for TF-IDF and PCA
- Use a Neural Network for prediction