Quora Question pairs

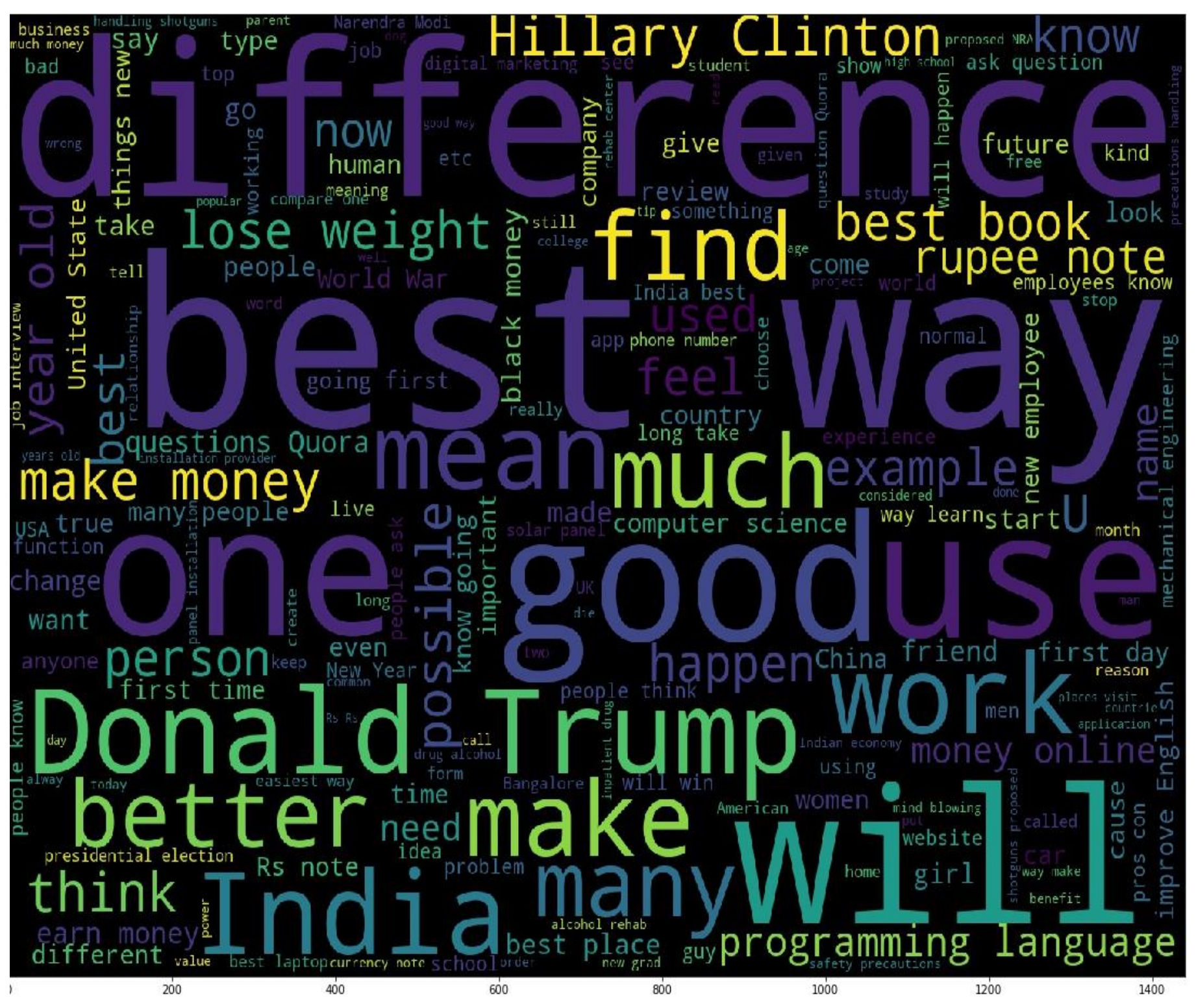
--Sentence Similarity Testing

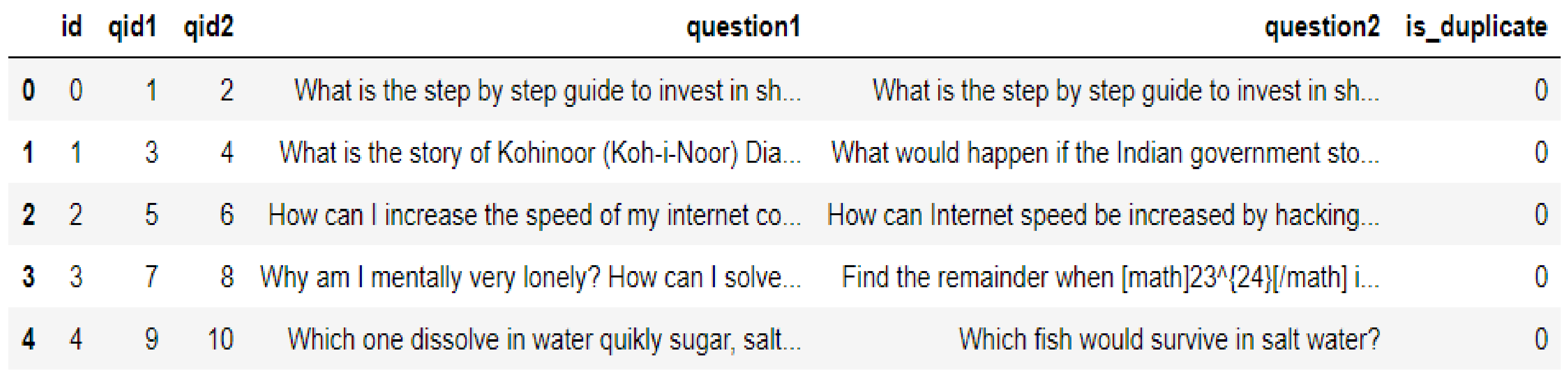
SFU
SIMON FRASER UNIVERSITY

Yifan Liu, Haosen Cheng
{yla416, haosenc}@sfu.ca

Background

Imagining you are asking a question on Quora, a quasi-forum website with over 100 million visitors a month. How does the system of Quora knows if this is a asked question? Auto detecting these duplicated question could be useful so people don't have to worry about finding them by hand. For example, to find an answer of a question, we can know if it's asked so we don't have to post it again. Here our goal is to identify which questions asked has already been asked before.

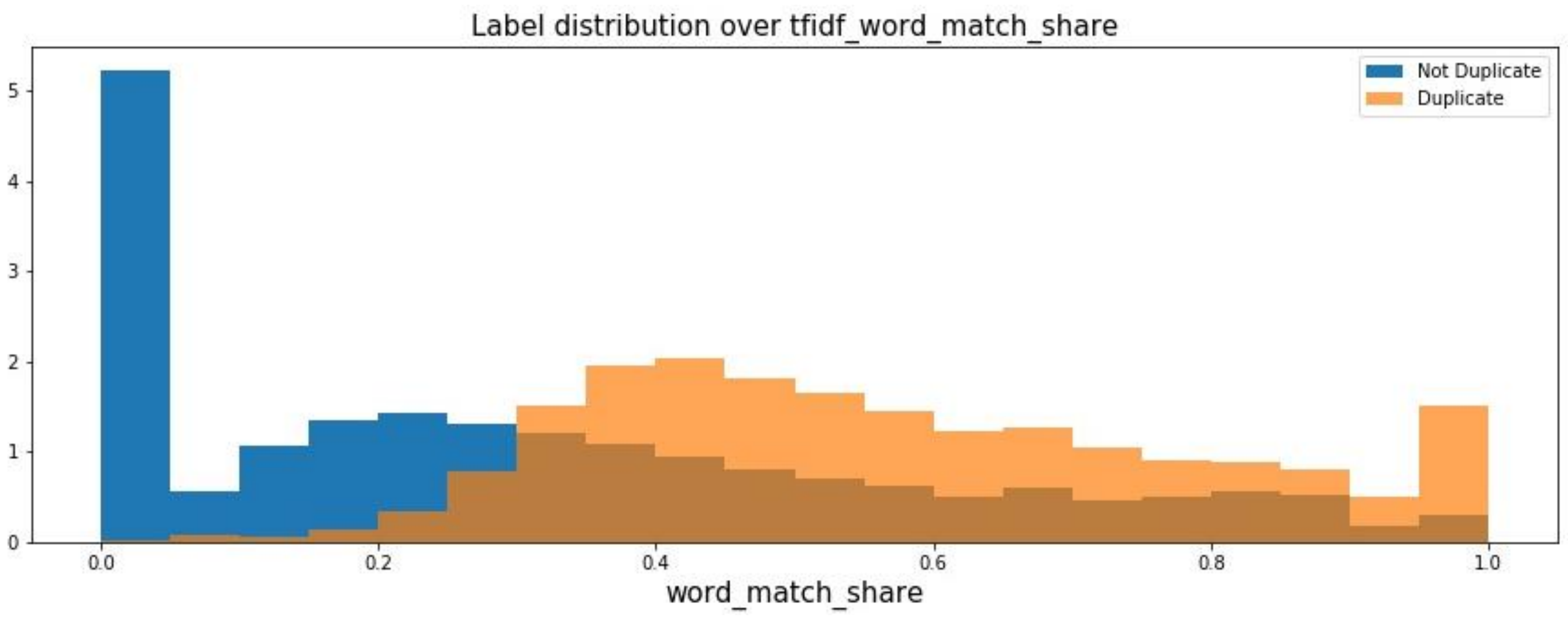




	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0

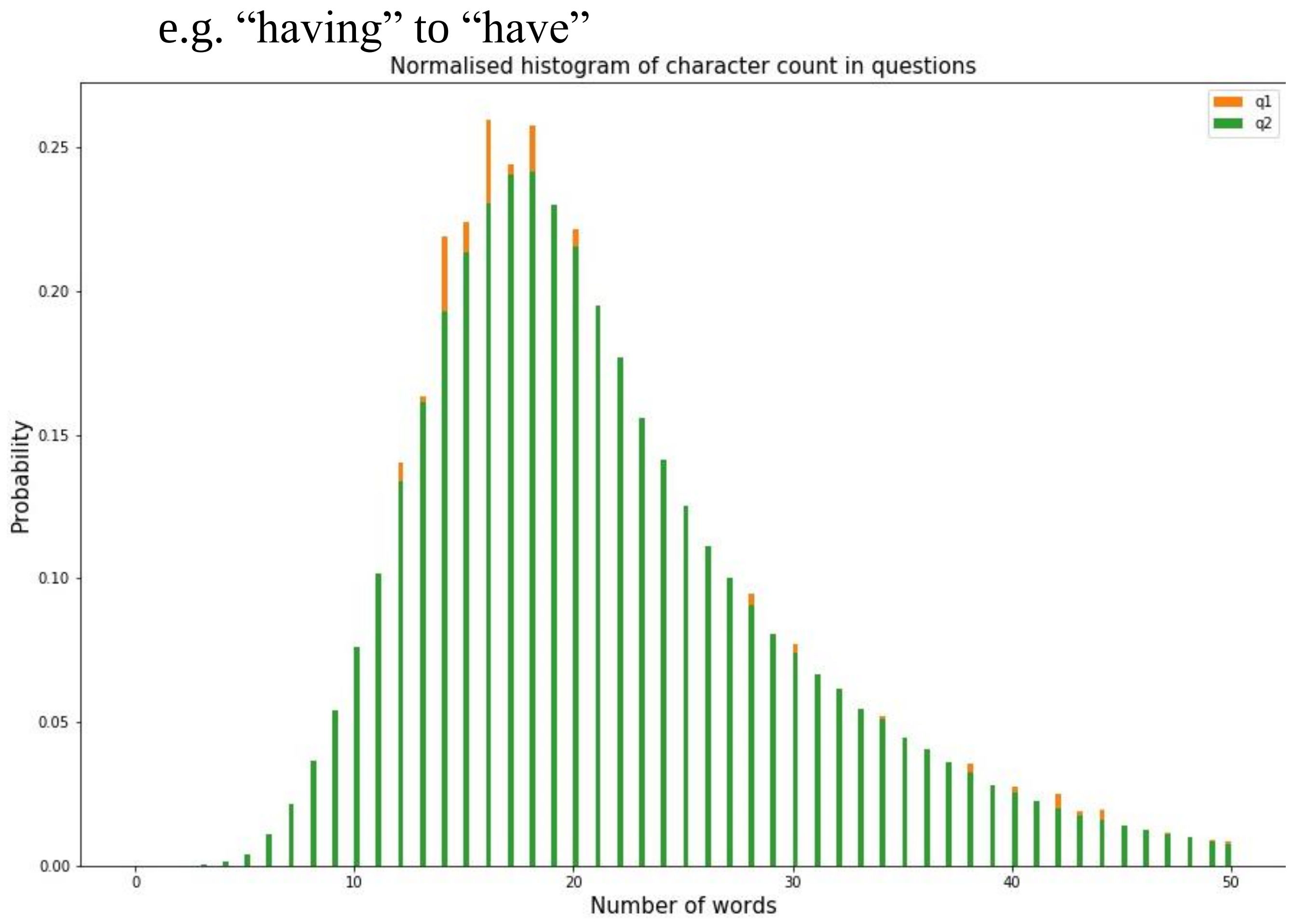
Methodology

- Pandas
  - Use pandas for basic data processing and storage
  - Provide basic statistic of train data
- Term frequency- inverse document frequency(TF-IDF)
  - Use to determine the importance of words
  - Fed into XGBoost for model training
- Word2vector
  - A pre-trained model by google
- NLTK
  - Use 2-gram and 3-gram features
  - Generate common length and common ratio of questions in train
- Other methods
  - Use plots to have a basic look of the data
    - Question length
  - Classify type of questions
  - Use pickle to save and load trained model
- XGBoost
  - A pre-implemented library to train model and make prediction on test data

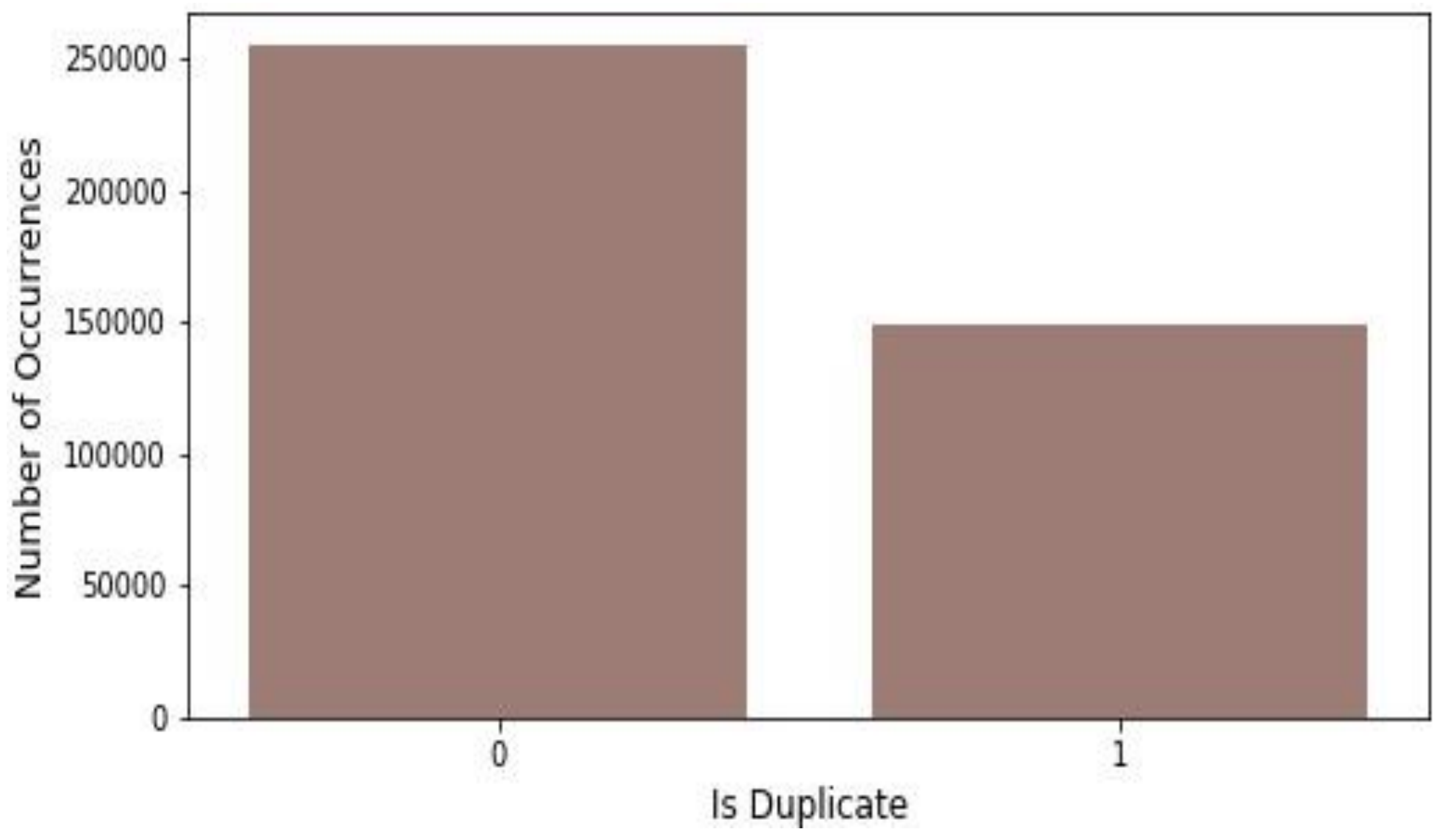


Preprocessing Data

- For the data cleaning, we filled empty entries with an empty string and removed duplicated rows.
- For easy computing of word count
  - We lowercased all letters
  - We changed all special forms like "I'm", "can't" to "I am", "cannot" etc.
  - Change proper Noun to it's relative Noun.
    - e.g. "IOS" to "operating system", "iPhone" to "phone"
- Deleted stop words using nltk library
- Used stemmer to change word format
  e.g. "having" to "have"



Compare the distribution of word count, we found that questions in train set and test have similar distribution.



Count of duplicates in training set

Conclusion

After some data cleaning and feature extraction, we increased degree of features of the train data. We then trained our model with XGBoost and made predictions on test data. By comparing of the results of different models we trained during this project, our best model have about 70% accuracy. Our future research will be pointing at deep learning and we are hoping to get a better accuracy at recognizing duplicate/similar questions.

Acknowledgements

https://www.kaggle.com/c/quora-question-pairs

https://www.kaggle.com/anokas/data-analysis-xgboost-starter-0-35460-lb

https://www.kaggle.com/currie32/the-importance-of-cleaning-text

http://blog.christianperone.com/2011/09/machine-learning-text-feature-extraction-tf-idf-part-i/

http://blog.christianperone.com/2011/10/machine-learning-text-feature-extraction-tf-idf-part-ii/

https://code.google.com/archive/p/word2vec/

http://xgboost.readthedocs.io/en/latest/