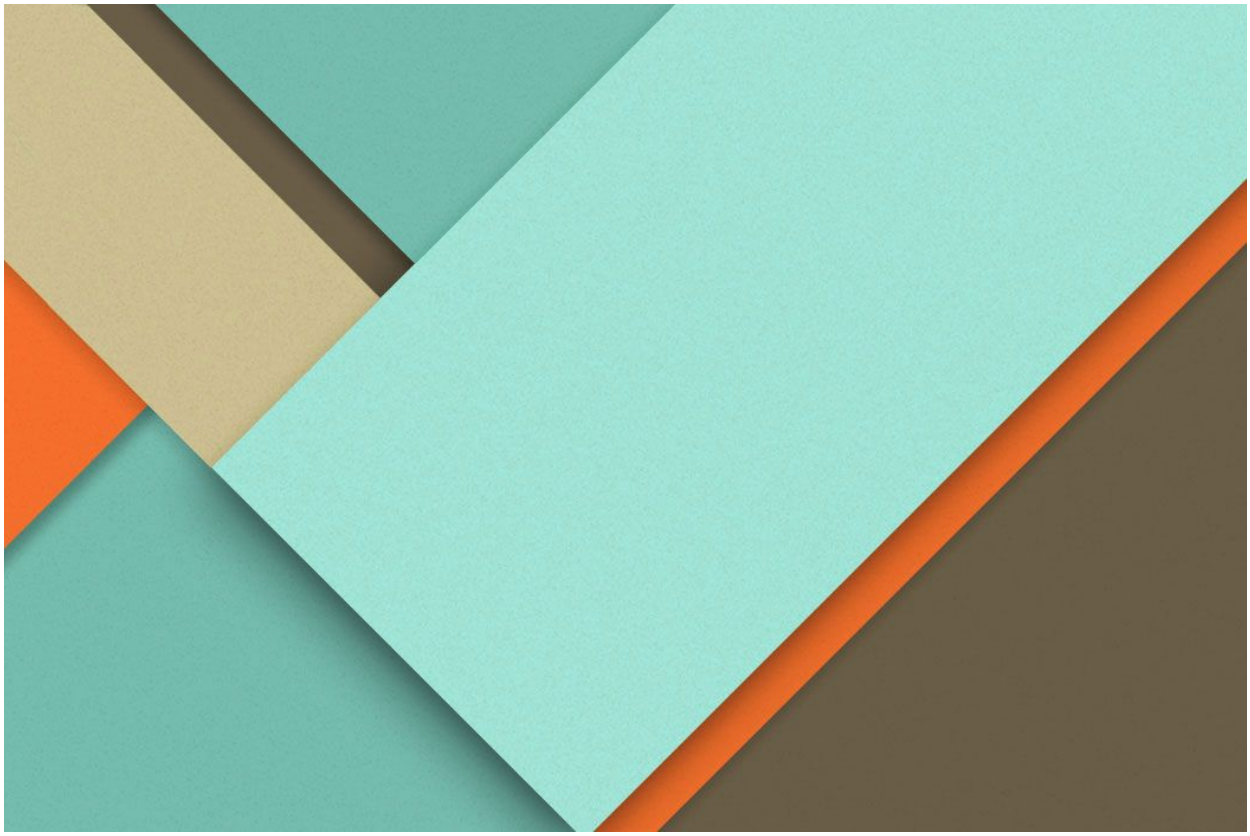


Semantic Question matching

Quora Question Pairs

1701CS36- Rahul Grover
1701CS19- Diksha Bansal

Data Visualization

Last two weeks, we visualised and analysed the training data set(provided by quora) available to us using python libraries:

1. **Numpy**: to do complicated mathematical calculations as numpy provides basic tools to compute with and manipulate the arrays.
2. **Pandas:** Pandas is one of those packages, and makes importing and analyzing data much easier. Pandas builds on packages like NumPy and matplotlib to give you a single, convenient, place to do most of your data analysis and visualization work.
3. **Matplotlib.pyplot :** matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

Results:

First Few rows of the training data set:

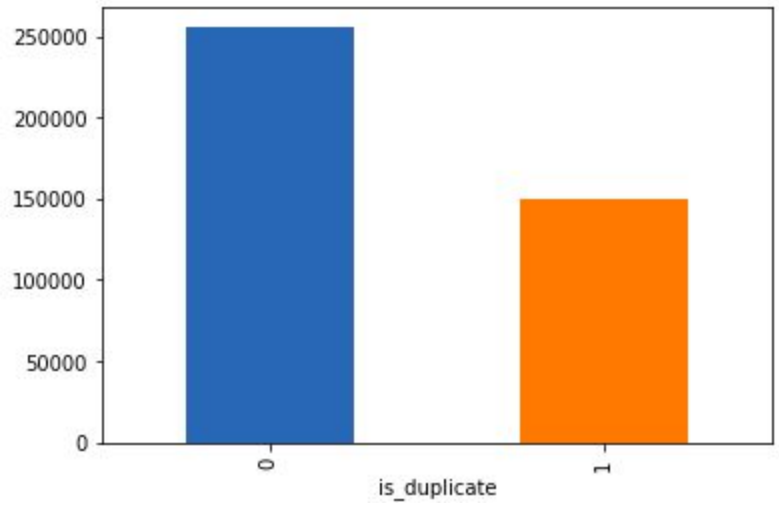
	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed 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

Training data information:

```
RangeIndex: 404290 entries, 0 to 404289
Data columns (total 6 columns):
id              404290 non-null int64
qid1            404290 non-null int64
qid2            404290 non-null int64
question1       404289 non-null object
question2       404288 non-null object
is_duplicate    404290 non-null int64
dtypes: int64(4), object(2)
memory usage: 18.5+ MB
```

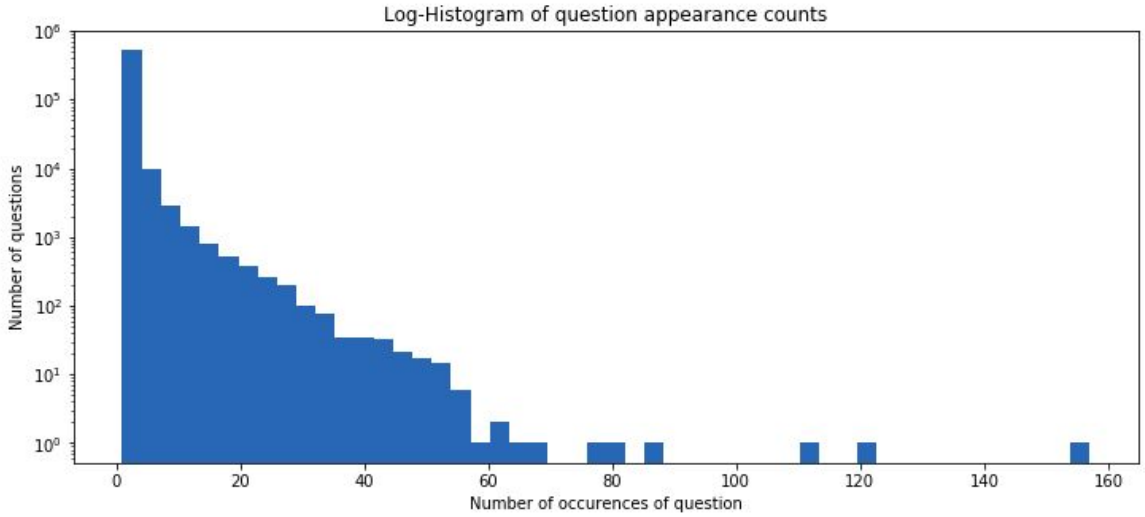
Total number of question pairs for training: 404290
Total number of unique questions in training set:537933
Number of questions that appear multiple times:111780

Questions with question marks: 99.87%
Questions with [math] tags: 0.12%
Questions with full stops: 6.31%
Questions with capitalized first letters: 99.81%
Questions with capital letters: 99.95%
Questions with numbers: 11.83%

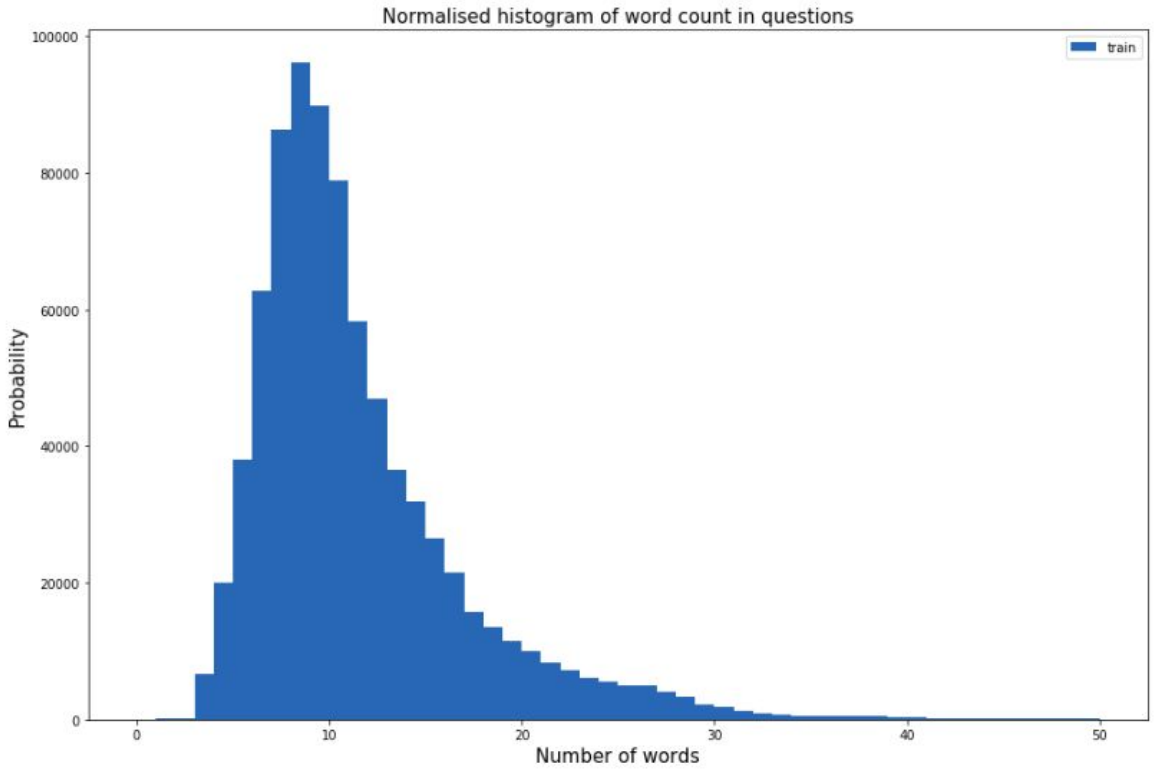


0- Number of non-duplicate question pairs.

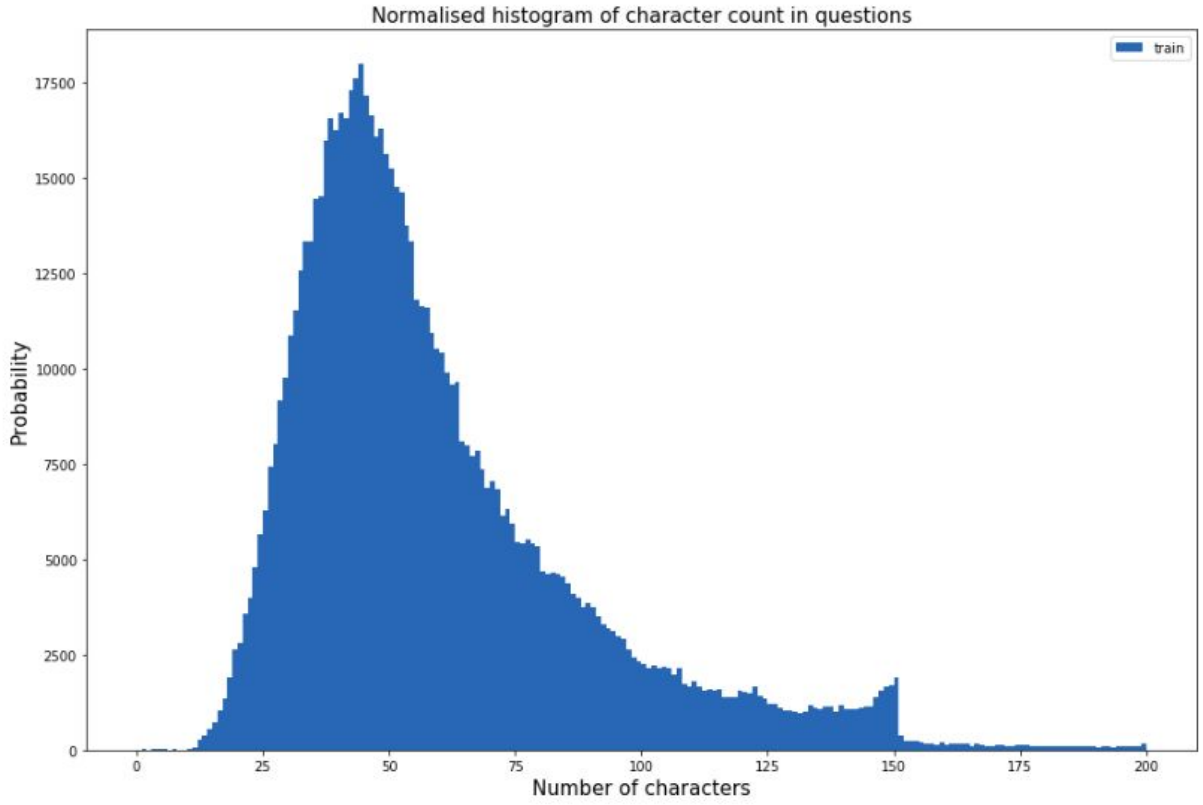
1- Number of duplicate question pairs.



Number of occurrences of question vs Number of question

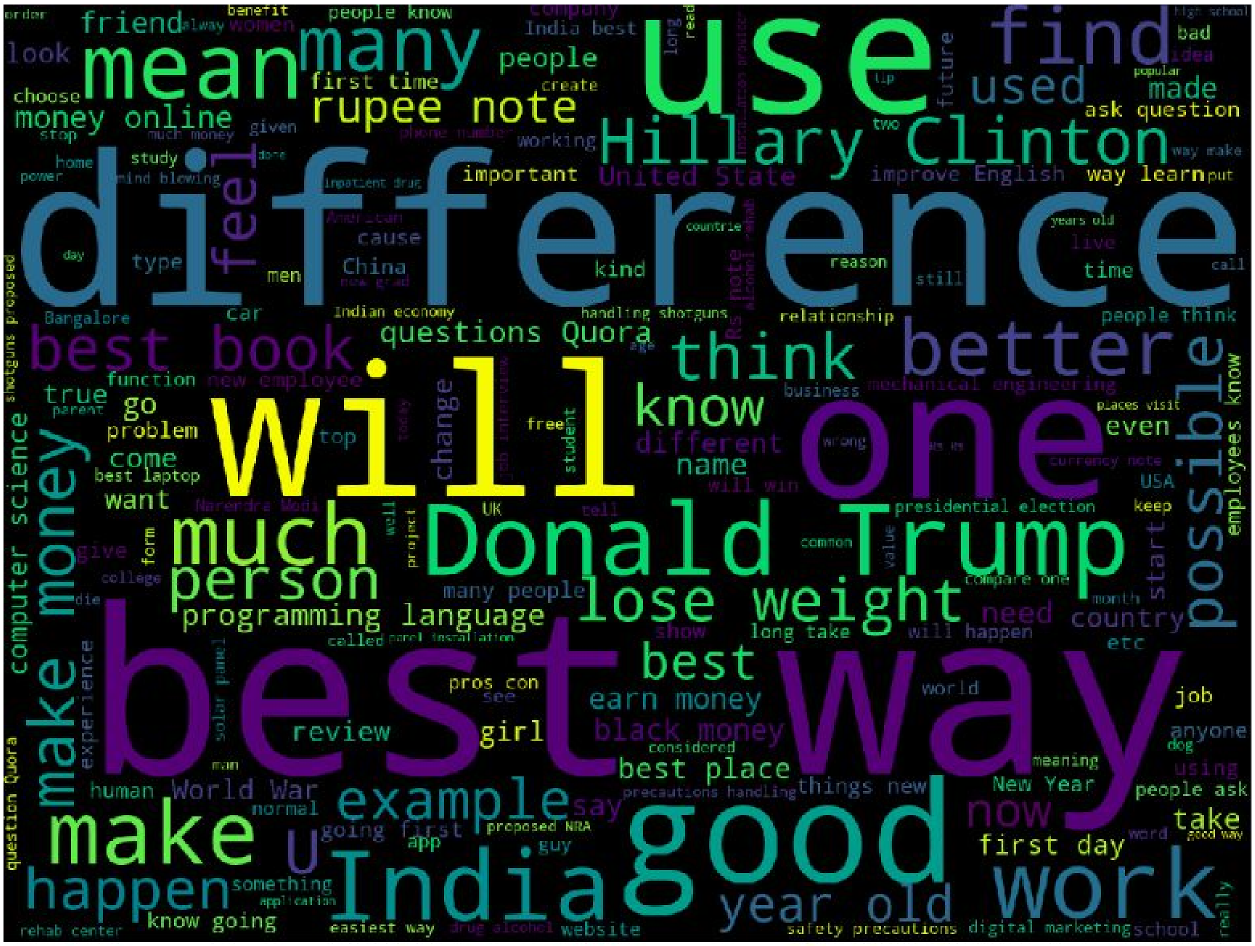


Number of words in a question vs number of questions



Number of characters in a question vs Number of questions

Word Cloud:



Conclusion:

We have completed visualization of data. In the coming week, we are planning to model using TF-IDF similarity and n-gram similarity.