[SENTENCE TO SENTENCE SEMANTIC SIMILARITY]

[Team ID : 22]

20191700280	1) ساره محمود كمال عبد الوهاب
20191700733	2) هبه الله علي سيد محمود
20191700735	3) هدى سامي محب عبد الحافظ
20191700734	4) هدى زين العابدين احمد ابراهيم
20191700089	5) أروى علاء الدين عبد الحميد

Sentence to Sentence semantic similarity

Introduction:

The project's main idea is to predict which of the provided pairs of questions contain two questions with the same meaning or not (duplicated or not).

Methodology:

➢ Preprocessing

- Change sentences to lowercase
- Replace abbreviations with their original using regular expression
- Tokenization
- Remove stop words
- lemmatization
- Replace nulls with an empty string

➢ Doc2Vec model

Doc2Vec model, as opposite to the Word2Vec model, is used to create a vectorized representation of a group of words taken collectively as a single unit. It doesn't only give the simple average of the words in the sentence.

It is preferred to use the doc2vec instance of word2vec when you have a set of sentences, not words

➢ Models

Using classification models to identify question pairs that have the same intent or meaning

- XGBoost Classifier
  - Train subset accuracy 0.8166909891414579
  - Test subset accuracy 0.8112493507135967

- AdaBoostst Classifier
  - Train subset accuracy 0.7998311855351357
  - Test subset accuracy 0.7998961141754681

Data Set Summary:

1-What is the data set used?

Quora Question Pairs in Kaggle (The train data set in the Link)

2- What is the summary of the dataset columns?

id → number of instances in data

qid1 → ids for first questions

qid2 → ids for second questions

- each question has one id and there are no two questions that have the same id

question1 → the text of the first questions

question2 → the text of the second questions

is_duplicate → classify the two questions are duplicate or not

- 0 → the two questions aren't duplicate or don't have the same meaning
- 1 → the two questions are duplicate or have the same meaning

cosine_similarity→ Cosine similarity measures document similarity in text analysis

cwc_min→ Get the common Tokens from the Question pair

cwc_max → Get the common Tokens from the Question pair

last_word_eq→ Last word of both questions is the same or not

first_word_eq→ First word of both questions is the same or not

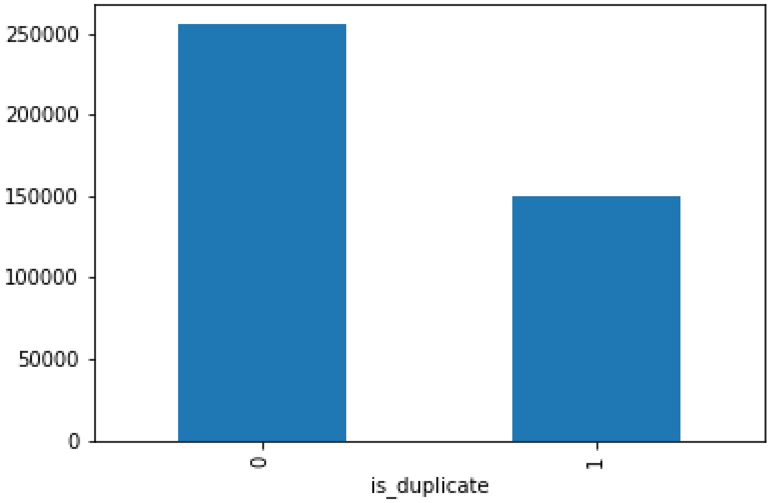
q1_ferq→ frequency of first questions in data

q2_ferq→ frequency of second questions in data

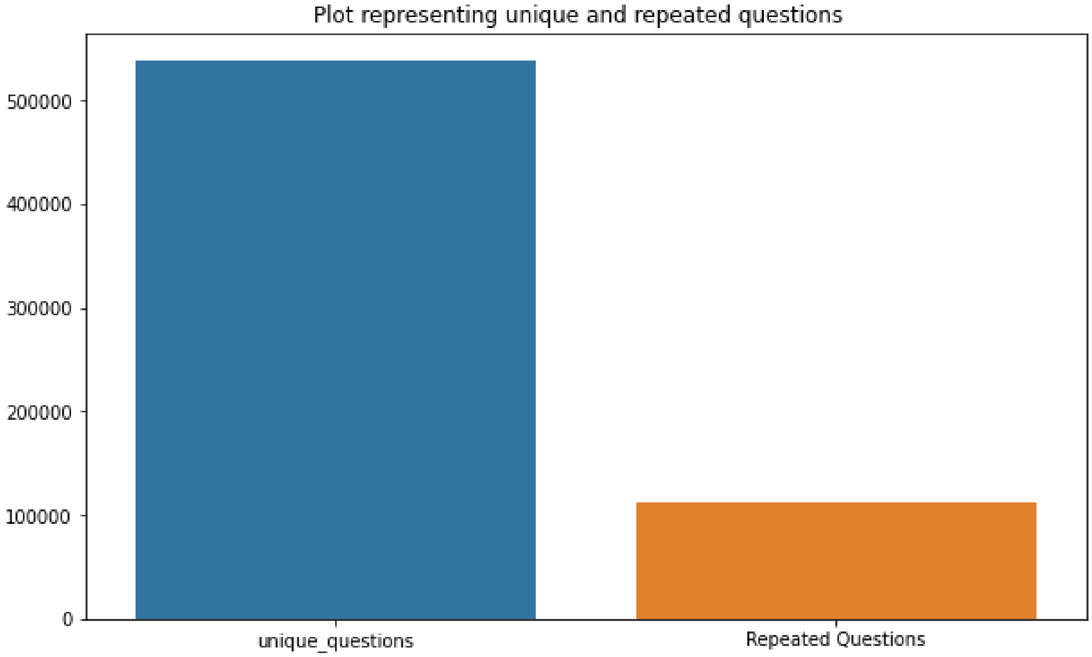
word_share→ word_common (number of common question pair) **/** word_total (length of question 1 + length of question 2)

3- Visualize the dataset statistics*/

- **bar chart for representing is_duplicated column**

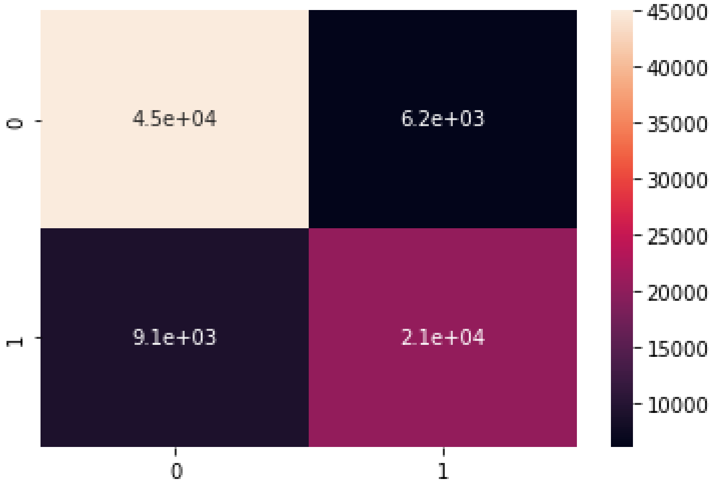


- **bar chart for representing unique and repeated questions**

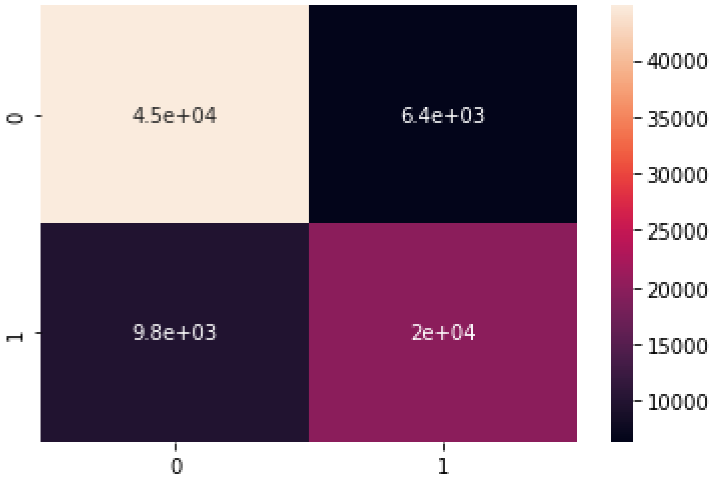


Results:

- **XGBoost Classifier**



- **AdaBoostst Classifier**



- **XGBoost model** has the highest accuracy where Training accuracy = 0. 8166909891414579 , Testing accuracy = 0. 8112493507135967