In [2]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import re
from bs4 import BeautifulSoup
```

In [3]:

```
df=pd.read_csv('train.csv')
df.head()
```

Out[3]:

	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

In [4]:

```
wrk_df=df.sample(25000,random_state=2)
wrk_df.head()
```

Out[4]:

	id	qid1	qid2	question1	question2	is_duplicate
398782	398782	496695	532029	What is the best marketing automation tool for...	What is the best marketing automation tool for...	1
115086	115086	187729	187730	I am poor but I want to invest. What should I do?	I am quite poor and I want to be very rich. Wh...	0
327711	327711	454161	454162	I am from India and live abroad. I met a guy f...	T.I.E.T to Thapar University to Thapar Univers...	0
367788	367788	498109	491396	Why do so many people in the U.S. hate the sou...	My boyfriend doesnt feel guilty when he hurts ...	0
151235	151235	237843	50930	Consequences of Bhopal gas tragedy?	What was the reason behind the Bhopal gas trag...	0

In [5]:

```
wrk_df['is_duplicate'].value_counts()
```

Out[5]:

```
0    15796
1     9204
Name: is_duplicate, dtype: int64
```

In [6]:

```
def preprocess(q):
    q=str(q).lower().strip()
    #replacing special character
    q = q.replace('%', ' percent')
    q = q.replace('$', ' dollar ')
    q = q.replace('₹', ' rupee ')
    q = q.replace('€', ' euro ')
    q = q.replace('@', ' at ')
    #observed in data anamosialy
    q = q.replace('[math]', '')
    #https://stackoverflow.com/a/19794953
    contractions = {
    "ain't": "am not",
    "aren't": "are not",
    "can't": "can not",
    "can't've": "can not have",
    "'cause": "because",
    "could've": "could have",
    "couldn't": "could not",
    "couldn't've": "could not have",
    "didn't": "did not",
    "doesn't": "does not",
    "don't": "do not",
    "hadn't": "had not",
    "hadn't've": "had not have",
    "hasn't": "has not",
    "haven't": "have not",
    "he'd": "he would",
    "he'd've": "he would have",
    "he'll": "he will",
    "he'll've": "he will have",
    "he's": "he is",
    "how'd": "how did",
    "how'd'y": "how do you",
    "how'll": "how will",
    "how's": "how is",
    "i'd": "i would",
    "i'd've": "i would have",
    "i'll": "i will",
    "i'll've": "i will have",
    "i'm": "i am",
    "i've": "i have",
    "isn't": "is not",
    "it'd": "it would",
    "it'd've": "it would have",
    "it'll": "it will",
    "it'll've": "it will have",
    "it's": "it is",
    "let's": "let us",
    "ma'am": "madam",
    "mayn't": "may not",
    "might've": "might have",
    "mightn't": "might not",
    "mightn't've": "might not have",
    "must've": "must have",
    "mustn't": "must not",
    "mustn't've": "must not have",
    "needn't": "need not",
    "needn't've": "need not have",
```

```
    "o'clock": "of the clock",
    "oughtn't": "ought not",
    "oughtn't've": "ought not have",
    "shan't": "shall not",
    "sha'n't": "shall not",
    "shan't've": "shall not have",
    "she'd": "she would",
    "she'd've": "she would have",
    "she'll": "she will",
    "she'll've": "she will have",
    "she's": "she is",
    "should've": "should have",
    "shouldn't": "should not",
    "shouldn't've": "should not have",
    "so've": "so have",
    "so's": "so as",
    "that'd": "that would",
    "that'd've": "that would have",
    "that's": "that is",
    "there'd": "there would",
    "there'd've": "there would have",
    "there's": "there is",
    "they'd": "they would",
    "they'd've": "they would have",
    "they'll": "they will",
    "they'll've": "they will have",
    "they're": "they are",
    "they've": "they have",
    "to've": "to have",
    "wasn't": "was not",
    "we'd": "we would",
    "we'd've": "we would have",
    "we'll": "we will",
    "we'll've": "we will have",
    "we're": "we are",
    "we've": "we have",
    "weren't": "were not",
    "what'll": "what will",
    "what'll've": "what will have",
    "what're": "what are",
    "what's": "what is",
    "what've": "what have",
    "when's": "when is",
    "when've": "when have",
    "where'd": "where did",
    "where's": "where is",
    "where've": "where have",
    "who'll": "who will",
    "who'll've": "who will have",
    "who's": "who is",
    "who've": "who have",
    "why's": "why is",
    "why've": "why have",
    "will've": "will have",
    "won't": "will not",
    "won't've": "will not have",
    "would've": "would have",
    "wouldn't": "would not",
    "wouldn't've": "would not have",
    "y'all": "you all",
    "y'all'd": "you all would",
```

```
    "y'all'd've": "you all would have",
    "y'all're": "you all are",
    "y'all've": "you all have",
    "you'd": "you would",
    "you'd've": "you would have",
    "you'll": "you will",
    "you'll've": "you will have",
    "you're": "you are",
    "you've": "you have"
    }
    q_decontracted = []

    for word in q.split():
        if word in contractions:
            word = contractions[word]

        q_decontracted.append(word)

    q = ' '.join(q_decontracted)
    q = q.replace("'ve", " have")
    q = q.replace("n't", " not")
    q = q.replace("'re", " are")
    q = q.replace("'ll", " will")

    # Removing HTML tags
    q = BeautifulSoup(q)
    q = q.get_text()

    # Remove punctuations
    pattern = re.compile('\W')
    q = re.sub(pattern, ' ', q).strip()


    return q
```

In [7]:

```
preprocess("I've already! wasn't <b>done</b>?")
```

Out[7]:

```
'i have already  was not done'
```

In [8]:

```
wrk_df['question1']=wrk_df['question1'].apply(preprocess)
wrk_df['question2']=wrk_df['question2'].apply(preprocess)
```

```
C:\Users\RESHAB\anaconda3\lib\site-packages\bs4\__init__.py:435: MarkupRes
emblesLocatorWarning: The input looks more like a filename than markup. Yo
u may want to open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
```

In [9]:

```
wrk_df.head()
```

Out[9]:

	id	qid1	qid2	question1	question2	is_duplicate
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1
115086	115086	187729	187730	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...	0
327711	327711	454161	454162	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...	0
367788	367788	498109	491396	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...	0
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0

In [10]:

```
wrk_df['q1_len'] = wrk_df['question1'].str.len()
wrk_df['q2_len'] = wrk_df['question2'].str.len()
```

In [11]:

```
wrk_df['q1_num_words'] = wrk_df['question1'].apply(lambda row: len(row.split(" ")))
wrk_df['q2_num_words'] = wrk_df['question2'].apply(lambda row: len(row.split(" ")))
wrk_df.head()
```

Out[11]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1	75	76	
115086	115086	187729	187730	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...	0	48	56	
327711	327711	454161	454162	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...	0	104	119	
367788	367788	498109	491396	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...	0	58	145	
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0	34	49	

In [12]:

```
def common_words(row):
    w1 = set(map(lambda word: word.lower().strip(), row['question1'].split(" ")))
    w2 = set(map(lambda word: word.lower().strip(), row['question2'].split(" ")))
    return len(w1 & w2)
wrk_df['word_common'] = wrk_df.apply(common_words, axis=1)
wrk_df.head()
```

Out[12]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1	75	76	
115086	115086	187729	187730	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...	0	48	56	
327711	327711	454161	454162	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...	0	104	119	
367788	367788	498109	491396	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...	0	58	145	
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0	34	49	

In [13]:

```
def total_words(row):
    w1 = set(map(lambda word: word.lower().strip(), row['question1'].split(" ")))
    w2 = set(map(lambda word: word.lower().strip(), row['question2'].split(" ")))
    return (len(w1) + len(w2))
wrk_df['word_total'] = wrk_df.apply(total_words, axis=1)
wrk_df.head()
```

Out[13]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1	75	76	
115086	115086	187729	187730	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...	0	48	56	
327711	327711	454161	454162	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...	0	104	119	
367788	367788	498109	491396	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...	0	58	145	
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0	34	49	

In [14]:

```
wrk_df['word_share'] = round(wrk_df['word_common']/wrk_df['word_total'],2)
wrk_df.head()
```

Out[14]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1	75	76	
115086	115086	187729	187730	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...	0	48	56	
327711	327711	454161	454162	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...	0	104	119	
367788	367788	498109	491396	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...	0	58	145	
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0	34	49	

In [15]:

```
from nltk.corpus import stopwords
```

In [16]:

```
def fetch_token_features(row):

    q1 = row['question1']
    q2 = row['question2']

    SAFE_DIV = 0.0001

    STOP_WORDS = stopwords.words("english")

    token_features = [0.0]*8

    # Converting the Sentence into Tokens:
    q1_tokens = q1.split()
    q2_tokens = q2.split()

    if len(q1_tokens) == 0 or len(q2_tokens) == 0:
        return token_features

    # Get the non-stopwords in Questions
    q1_words = set([word for word in q1_tokens if word not in STOP_WORDS])
    q2_words = set([word for word in q2_tokens if word not in STOP_WORDS])

    #Get the stopwords in Questions
    q1_stops = set([word for word in q1_tokens if word in STOP_WORDS])
    q2_stops = set([word for word in q2_tokens if word in STOP_WORDS])

    # Get the common non-stopwords from Question pair
    common_word_count = len(q1_words.intersection(q2_words))

    # Get the common stopwords from Question pair
    common_stop_count = len(q1_stops.intersection(q2_stops))

    # Get the common Tokens from Question pair
    common_token_count = len(set(q1_tokens).intersection(set(q2_tokens)))


    token_features[0] = common_word_count / (min(len(q1_words), len(q2_words)) + SAFE_DI
    token_features[1] = common_word_count / (max(len(q1_words), len(q2_words)) + SAFE_DI
    token_features[2] = common_stop_count / (min(len(q1_stops), len(q2_stops)) + SAFE_DI
    token_features[3] = common_stop_count / (max(len(q1_stops), len(q2_stops)) + SAFE_DI
    token_features[4] = common_token_count / (min(len(q1_tokens), len(q2_tokens)) + SAFE
    token_features[5] = common_token_count / (max(len(q1_tokens), len(q2_tokens)) + SAFE

    # Last word of both question is same or not
    token_features[6] = int(q1_tokens[-1] == q2_tokens[-1])

    # First word of both question is same or not
    token_features[7] = int(q1_tokens[0] == q2_tokens[0])

    return token_features
```

In [17]:

```
token_features = wrk_df.apply(fetch_token_features, axis=1)

wrk_df["cwc_min"]       = list(map(lambda x: x[0], token_features))
wrk_df["cwc_max"]       = list(map(lambda x: x[1], token_features))
wrk_df["csc_min"]       = list(map(lambda x: x[2], token_features))
wrk_df["csc_max"]       = list(map(lambda x: x[3], token_features))
wrk_df["ctc_min"]       = list(map(lambda x: x[4], token_features))
wrk_df["ctc_max"]       = list(map(lambda x: x[5], token_features))
wrk_df["last_word_eq"]  = list(map(lambda x: x[6], token_features))
wrk_df["first_word_eq"] = list(map(lambda x: x[7], token_features))
```

In [18]:

```
wrk_df.head()
```

Out[18]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1	75	76	
115086	115086	187729	187730	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...	0	48	56	
327711	327711	454161	454162	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...	0	104	119	
367788	367788	498109	491396	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...	0	58	145	
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0	34	49	

5 rows × 21 columns

In [19]:

```
pip install distance
```

```
Requirement already satisfied: distance in c:\users\reshab\anaconda3\lib\s
ite-packages (0.1.3)
Note: you may need to restart the kernel to use updated packages.
```

In [20]:

```
import distance

def fetch_length_features(row):

    q1 = row['question1']
    q2 = row['question2']

    length_features = [0.0]*3

    # Converting the Sentence into Tokens:
    q1_tokens = q1.split()
    q2_tokens = q2.split()

    if len(q1_tokens) == 0 or len(q2_tokens) == 0:
        return length_features

    # Absolute length features
    length_features[0] = abs(len(q1_tokens) - len(q2_tokens))

    #Average Token Length of both Questions
    length_features[1] = (len(q1_tokens) + len(q2_tokens))/2

    strs = list(distance.lcsubstrings(q1, q2))
    length_features[2] = len(strs[0]) / (min(len(q1), len(q2)) + 1)

    return length_features
```

In [21]:

```
length_features = wrk_df.apply(fetch_length_features, axis=1)

wrk_df['abs_len_diff'] = list(map(lambda x: x[0], length_features))
wrk_df['mean_len'] = list(map(lambda x: x[1], length_features))
wrk_df['longest_substr_ratio'] = list(map(lambda x: x[2], length_features))
```

In [22]:

```
wrk_df.head()
```

Out[22]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1	75	76	
115086	115086	187729	187730	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...	0	48	56	
327711	327711	454161	454162	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...	0	104	119	
367788	367788	498109	491396	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...	0	58	145	
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0	34	49	

5 rows × 24 columns

In [23]:

```
pip install fuzzywuzzy
```

```
Requirement already satisfied: fuzzywuzzy in c:\users\reshab\anaconda3\lib
\site-packages (0.18.0)
Note: you may need to restart the kernel to use updated packages.
```

In [24]:

```
# Fuzzy Features
from fuzzywuzzy import fuzz

def fetch_fuzzy_features(row):

    q1 = row['question1']
    q2 = row['question2']

    fuzzy_features = [0.0]*4

    # fuzz_ratio
    fuzzy_features[0] = fuzz.QRatio(q1, q2)

    # fuzz_partial_ratio
    fuzzy_features[1] = fuzz.partial_ratio(q1, q2)

    # token_sort_ratio
    fuzzy_features[2] = fuzz.token_sort_ratio(q1, q2)

    # token_set_ratio
    fuzzy_features[3] = fuzz.token_set_ratio(q1, q2)

    return fuzzy_features
```

```
C:\Users\RESHAB\anaconda3\lib\site-packages\fuzzywuzzy\fuzz.py:11: UserWar
ning: Using slow pure-python SequenceMatcher. Install python-Levenshtein t
o remove this warning
  warnings.warn('Using slow pure-python SequenceMatcher. Install python-Le
venshtein to remove this warning')
```

In [25]:

```
fuzzy_features = wrk_df.apply(fetch_fuzzy_features, axis=1)

# Creating new feature columns for fuzzy features
wrk_df['fuzz_ratio'] = list(map(lambda x: x[0], fuzzy_features))
wrk_df['fuzz_partial_ratio'] = list(map(lambda x: x[1], fuzzy_features))
wrk_df['token_sort_ratio'] = list(map(lambda x: x[2], fuzzy_features))
wrk_df['token_set_ratio'] = list(map(lambda x: x[3], fuzzy_features))
```

In [26]:

```
print(wrk_df.shape)
wrk_df.head()
```

(25000, 28)

Out[26]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1	75	76	
115086	115086	187729	187730	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...	0	48	56	
327711	327711	454161	454162	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...	0	104	119	
367788	367788	498109	491396	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...	0	58	145	
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0	34	49	

5 rows × 28 columns

In [27]:

```
wrk_df.isnull().sum()
```

Out[27]:

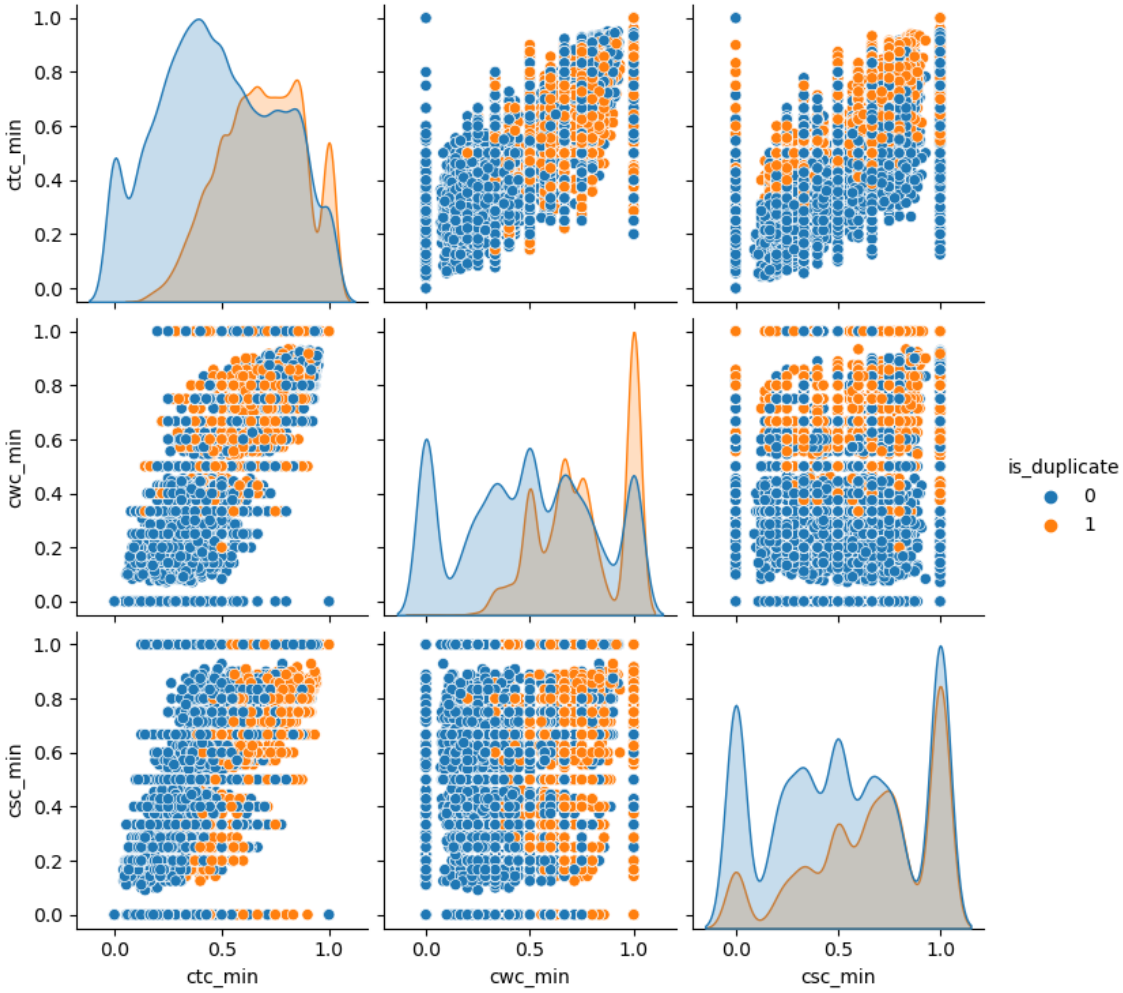
```
id                    0
qid1                  0
qid2                  0
question1             0
question2             0
is_duplicate          0
q1_len                0
q2_len                0
q1_num_words          0
q2_num_words          0
word_common           0
word_total            0
word_share            0
cwc_min               0
cwc_max               0
csc_min               0
csc_max               0
ctc_min               0
ctc_max               0
last_word_eq          0
first_word_eq         0
abs_len_diff          0
mean_len              0
longest_substr_ratio  0
fuzz_ratio            0
fuzz_partial_ratio    0
token_sort_ratio      0
token_set_ratio       0
dtype: int64
```

In [28]:

```
sns.pairplot(wrk_df[['ctc_min', 'cwc_min', 'csc_min', 'is_duplicate']],hue='is_duplicate
```

Out[28]:

```
<seaborn.axisgrid.PairGrid at 0x2c8804c1ae0>
```

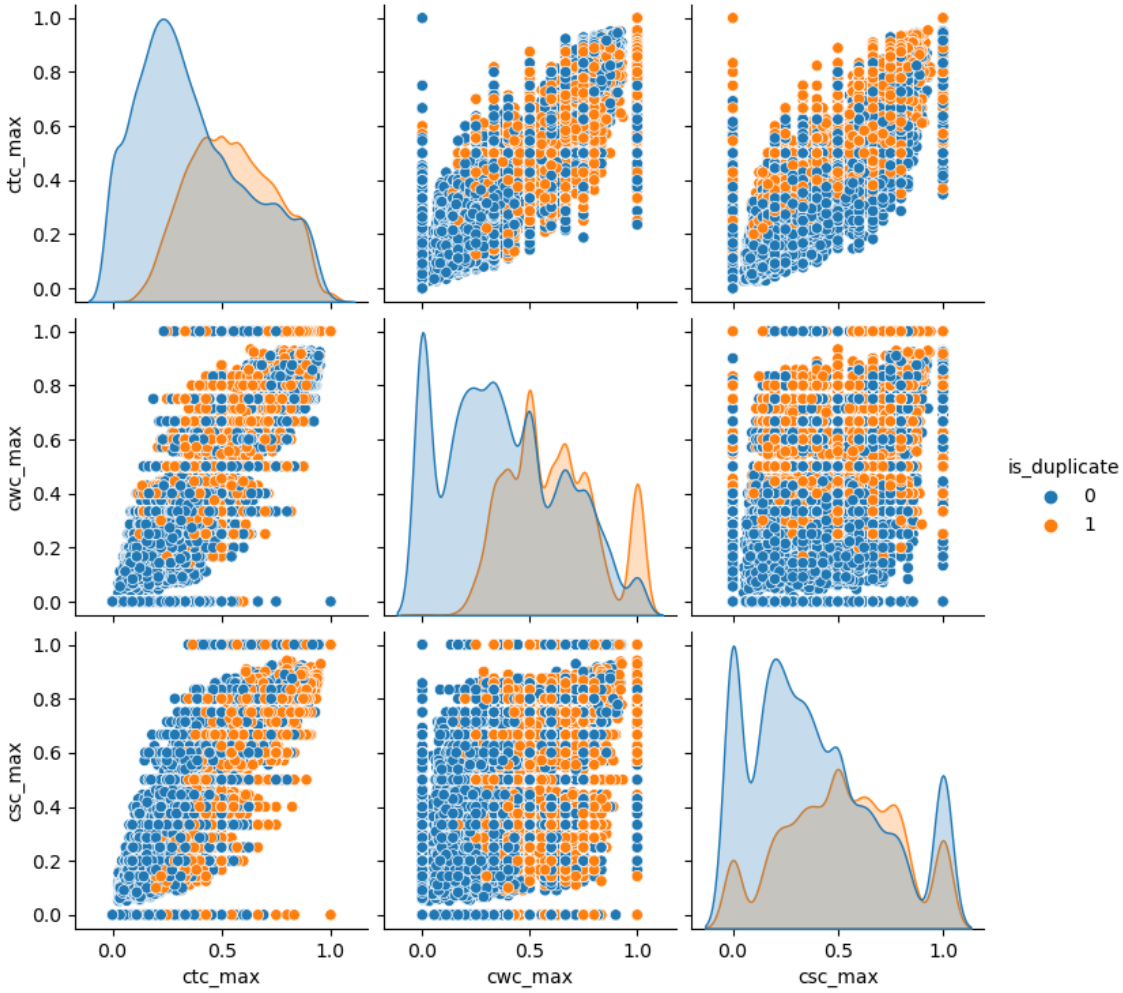


In [29]:

```
sns.pairplot(wrk_df[['ctc_max', 'cwc_max', 'csc_max', 'is_duplicate']],hue='is_duplicate
```

Out[29]:

```
<seaborn.axisgrid.PairGrid at 0x2c8821f61a0>
```

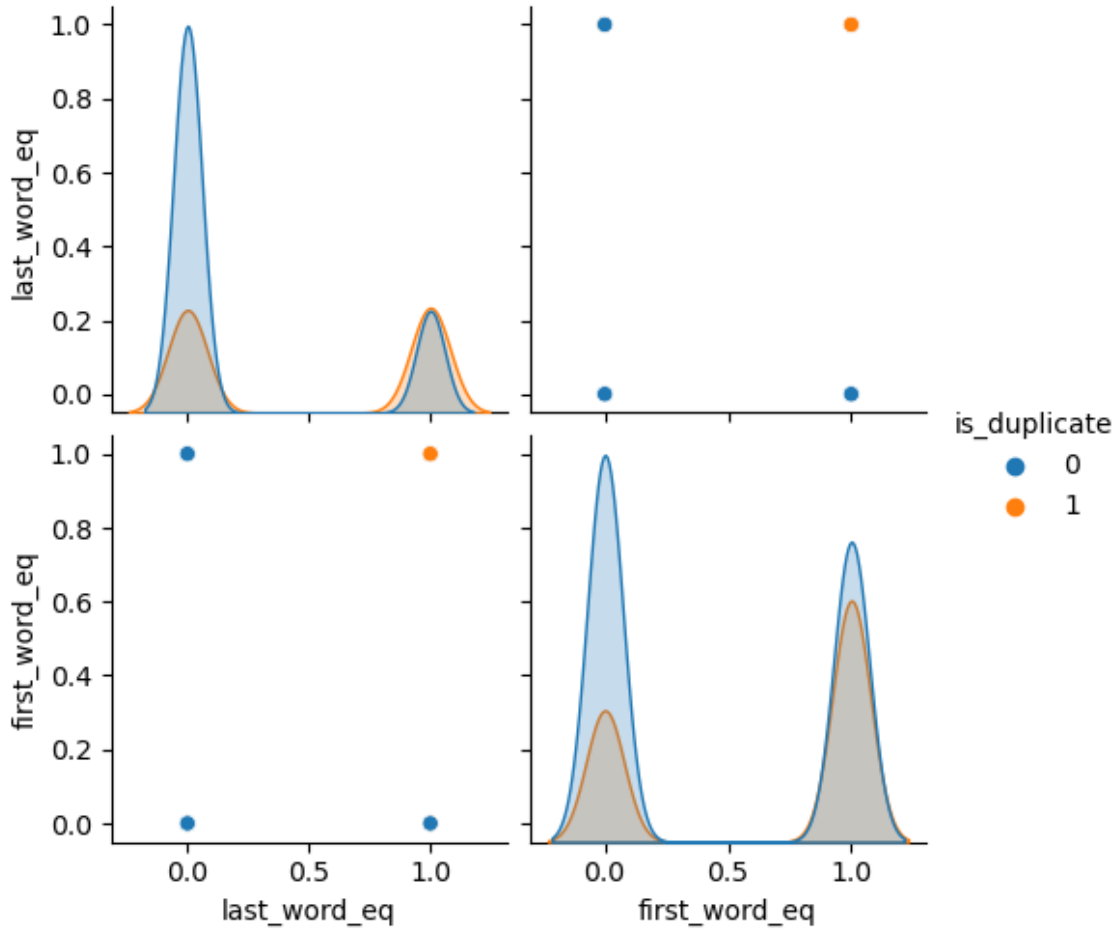


In [30]:

```
sns.pairplot(wrk_df[['last_word_eq', 'first_word_eq', 'is_duplicate']],hue='is_duplicate
```

Out[30]:

```
<seaborn.axisgrid.PairGrid at 0x2c88966b310>
```

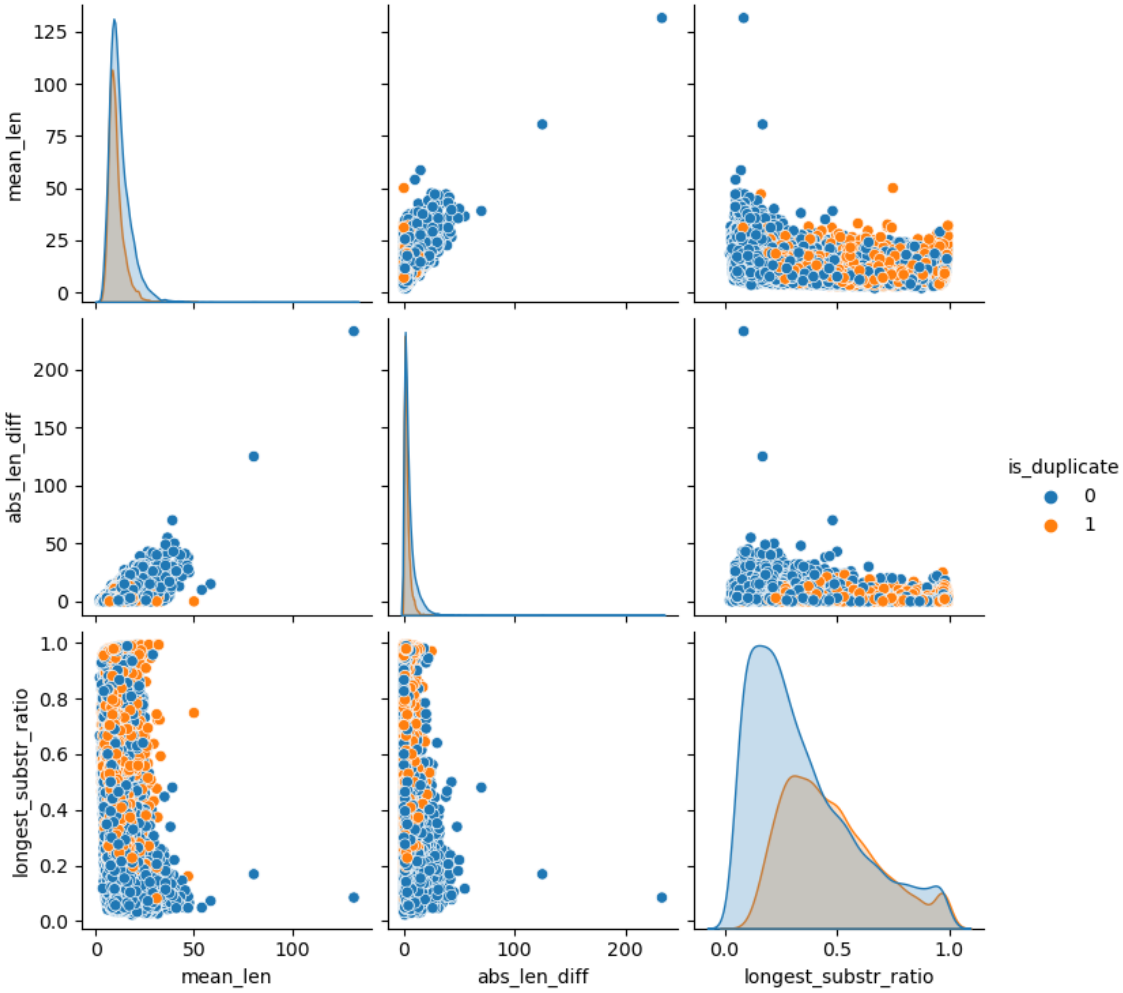


In [31]:

```
sns.pairplot(wrk_df[['mean_len', 'abs_len_diff','longest_substr_ratio', 'is_duplicate']]
```

Out[31]:

```
<seaborn.axisgrid.PairGrid at 0x2c88a8611e0>
```

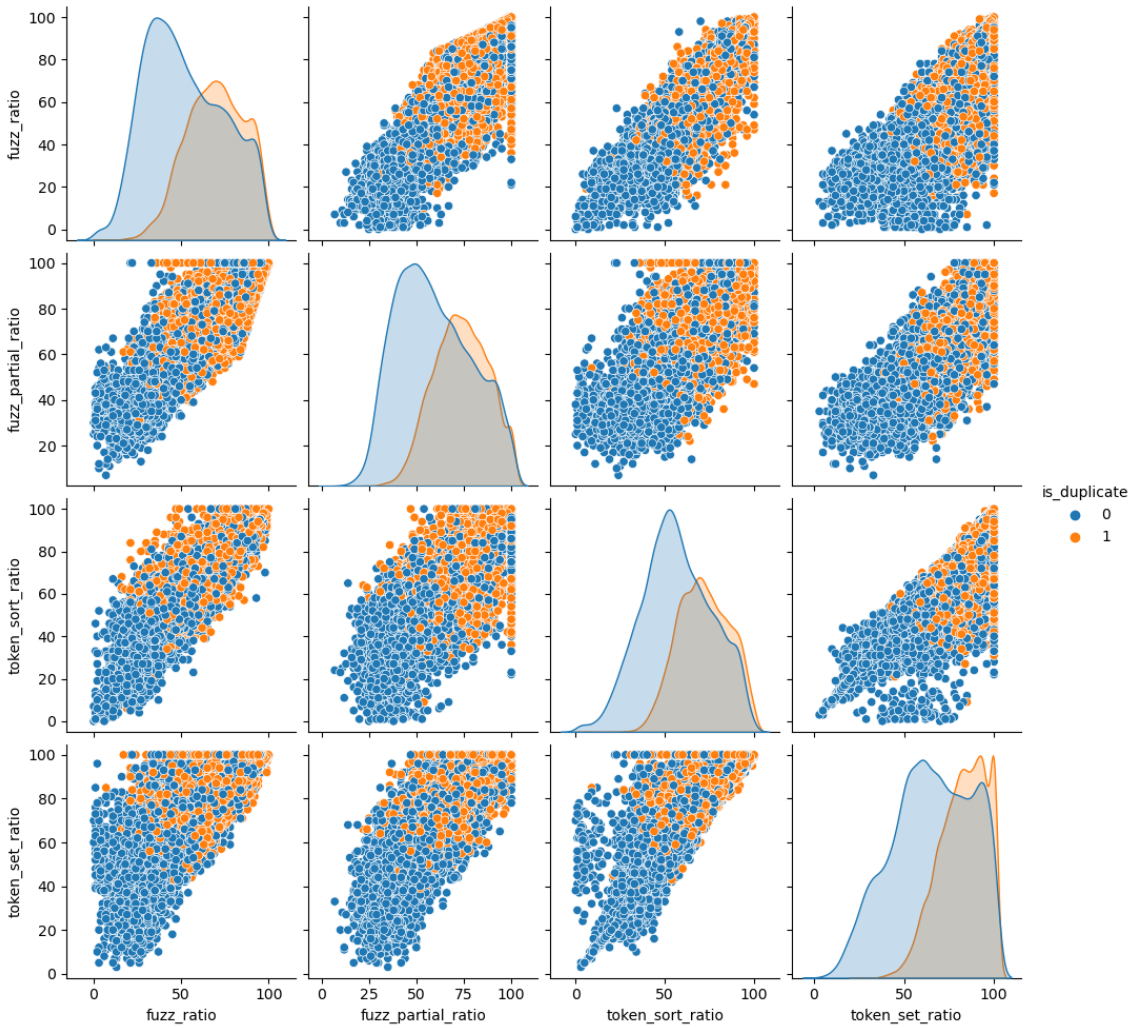


In [32]:

```
sns.pairplot(wrk_df[['fuzz_ratio', 'fuzz_partial_ratio','token_sort_ratio','token_set_ra
```

Out[32]:

```
<seaborn.axisgrid.PairGrid at 0x2c88b2cece0>
```



In [33]:

```
wrk_df.head()
```

Out[33]:

	id	qid1	qid2	question1	question2	is_duplicate	q1_len	q2_len	q1_
398782	398782	496695	532029	what is the best marketing automation tool for...	what is the best marketing automation tool for...	1	75	76	
115086	115086	187729	187730	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...	0	48	56	
327711	327711	454161	454162	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...	0	104	119	
367788	367788	498109	491396	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...	0	58	145	
151235	151235	237843	50930	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy	0	34	49	

5 rows × 28 columns

In [34]:

```
from sklearn.preprocessing import MinMaxScaler

X = MinMaxScaler().fit_transform(wrk_df[['cwc_min', 'cwc_max', 'csc_min', 'csc_max' , 'c
y =wrk_df['is_duplicate'].values
```

In [35]:

```
ques_df = wrk_df[['question1','question2']]
ques_df.head()
```

Out[35]:

	question1	question2
398782	what is the best marketing automation tool for...	what is the best marketing automation tool for...
115086	i am poor but i want to invest what should i do	i am quite poor and i want to be very rich wh...
327711	i am from india and live abroad i met a guy f...	t i e t to thapar university to thapar univers...
367788	why do so many people in the u s hate the sou...	my boyfriend doesnt feel guilty when he hurts ...
151235	consequences of bhopal gas tragedy	what was the reason behind the bhopal gas tragedy

In [36]:

```
final_df = wrk_df.drop(columns=['id','qid1','qid2','question1','question2'])
print(final_df.shape)
final_df.head()
```

```
(25000, 23)
```

Out[36]:

	is_duplicate	q1_len	q2_len	q1_num_words	q2_num_words	word_common	word_t
398782	1	75	76	13	13	12	
115086	0	48	56	13	16	8	
327711	0	104	119	28	21	4	
367788	0	58	145	14	32	1	
151235	0	34	49	5	9	3	

5 rows × 23 columns

In [37]:

```
from sklearn.feature_extraction.text import CountVectorizer
# merge texts
questions = list(ques_df['question1']) + list(ques_df['question2'])

cv = CountVectorizer(max_features=3000)
q1_arr, q2_arr = np.vsplit(cv.fit_transform(questions).toarray(),2)
```

In [38]:

```
temp_df1 = pd.DataFrame(q1_arr, index= ques_df.index)
temp_df2 = pd.DataFrame(q2_arr, index= ques_df.index)
temp_df = pd.concat([temp_df1, temp_df2], axis=1)
temp_df.shape
```

Out[38]:

```
(25000, 6000)
```

In [39]:

```
final_df = pd.concat([final_df, temp_df], axis=1)
print(final_df.shape)
final_df.head()
```

```
(25000, 6023)
```

Out[39]:

	is_duplicate	q1_len	q2_len	q1_num_words	q2_num_words	word_common	word_t
398782	1	75	76	13	13	12	
115086	0	48	56	13	16	8	
327711	0	104	119	28	21	4	
367788	0	58	145	14	32	1	
151235	0	34	49	5	9	3	

5 rows × 6023 columns

In [40]:

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(final_df.iloc[:,1:].values,final_df.ilo
```

In [41]:

```
# from sklearn.ensemble import RandomForestClassifier
 from sklearn.metrics import accuracy_score
# rf = RandomForestClassifier()
# rf.fit(X_train,y_train)
# y_pred = rf.predict(X_test)
# accuracy_score(y_test,y_pred)
```

In [42]:

```
# from xgboost import XGBClassifier
# xgb = XGBClassifier()
# xgb.fit(X_train,y_train)
# y_pred1 = xgb.predict(X_test)
# accuracy_score(y_test,y_pred1)
```

In [ ]:

```
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', C = 0.1, gamma = 0.1)
classifier.fit(X_train, y_train)
y_pred = rf.predict(X_test)
accuracy_score(y_test,y_pred)
```