```
import pandas as pd
file_path = "quora_questions.csv"
data_set = pd.read_csv(file_path)
data_set.head()
```

	Question
0	What is the step by step guide to invest in sh...
1	What is the story of Kohinoor (Koh-i-Noor) Dia...
2	How can I increase the speed of my internet co...
3	Why am I mentally very lonely? How can I solve...
4	Which one dissolve in water quikly sugar, salt...

```
import nltk
import string
import re


from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer



#Removing special characters and convert to lowercase
def clean_text(text):
    return re.sub(r'[^a-zA-Z\s]', '', text).lower()

data_set['Cleaned_Questions'] = data_set['Question'].apply(clean_text)
data_set
```

	Question	Cleaned_Questions
0	What is the step by step guide to invest in sh...	what is the step by step guide to invest in sh...
1	What is the story of Kohinoor (Koh-i-Noor) Dia...	what is the story of kohinoor kohinoor diamond
2	How can I increase the speed of my internet co...	how can i increase the speed of my internet co...
3	Why am I mentally very lonely? How can I solve...	why am i mentally very lonely how can i solve it
4	Which one dissolve in water quikly sugar, salt...	which one dissolve in water quikly sugar salt ...
...	...	...
404284	How many keywords are there in the Racket prog...	how many keywords are there in the racket prog...
404285	Do you believe there is life after death?	do you believe there is life after death
404286	What is one coin?	what is one coin
404287	What is the approx annual cost of living while...	what is the approx annual cost of living while...
404288	What is like to have sex with cousin?	what is like to have sex with cousin

404289 rows × 2 columns

```
# Removing stopwords
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

def stopwords_removal(text):
    words = text.split()
    filtered_words = []
    for word in words:
        if word not in stop_words:
            filtered_words.append(word)
    return ' '.join(filtered_words)

data_set['Cleaned_Questions'] = data_set['Cleaned_Questions']. apply(stopwords_removal)
data_set
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

	Question	Cleaned_Questions
0	What is the step by step guide to invest in sh...	step step guide invest share market india
1	What is the story of Kohinoor (Koh-i-Noor) Dia...	story kohinoor kohinoor diamond
2	How can I increase the speed of my internet co...	increase speed internet connection using vpn
3	Why am I mentally very lonely? How can I solve...	mentally lonely solve
4	Which one dissolve in water quikly sugar, salt...	one dissolve water quikly sugar salt methane c...
...	...	...
404284	How many keywords are there in the Racket prog...	many keywords racket programming language late...
404285	Do you believe there is life after death?	believe life death
404286	What is one coin?	one coin
404287	What is the approx annual cost of living while...	approx annual cost living studying uic chicago...
404288	What is like to have sex with cousin?	like sex cousin

404289 rows × 2 columns

```
#Tokenization
nltk.download('punkt')
data_set['tokenized_words'] = data_set['Cleaned_Questions'].apply(word_tokenize)
data_set.head()
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

	Question	Cleaned_Questions	tokenized_words
0	What is the step by step guide to invest in sh...	step step guide invest share market india	[step, step, guide, invest, share, market, india]
1	What is the story of Kohinoor (Koh-i-Noor) Dia...	story kohinoor kohinoor diamond	[story, kohinoor, kohinoor, diamond]
2	How can I increase the speed of my internet co...	increase speed internet connection using vpn	[increase, speed, internet, connection, using,...
3	Why am I mentally very lonely? How can I solve...	mentally lonely solve	[mentally, lonely, solve]
4	Which one dissolve in water quikly sugar, salt...	one dissolve water quikly sugar salt methane c...	[one, dissolve, water, quikly, sugar, salt, me...

Start coding or generate with AI.