explore_tokens_per_sentence

September 11, 2022

[ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from transformers import AutoTokenizer

tokenizer = AutoTokenizer.from_pretrained("distilbert-base-uncased")
```

[ ]:
```
train_df = pd.read_csv("../../data/raw/train.csv")
test_df = pd.read_csv("../../data/raw/test.csv")
```

[ ]:
```
train_df.head()
```

[ ]:
```
        text_id                                          full_text  cohesion  \
0  0016926B079C  I think that students would benefit from learn…       3.5
1  0022683E9EA5  When a problem is a change you have to let it …       2.5
2  00299B378633  Dear, Principal\n\nIf u change the school poli…       3.0
3  003885A45F42  The best time in life is when you become yours…       4.5
4  0049B1DF5CCC  Small act of kindness can impact in other peop…       2.5

   syntax  vocabulary  phraseology  grammar  conventions
0     3.5         3.0          3.0      4.0          3.0
1     2.5         3.0          2.0      2.0          2.5
2     3.5         3.0          3.0      3.0          2.5
3     4.5         4.5          4.5      4.0          5.0
4     3.0         3.0          3.0      2.5          2.5
```

[ ]:
```

```

[ ]:
```
from modules.gec.preprocessing import preprocessing_pipeline
```

[ ]:
```
target_colnames = ["cohesion", "syntax", "vocabulary", "phraseology",␣
 ↪"grammar", "conventions"]

sentence_train_df = pd.DataFrame()
cnt = 0
for rec in train_df.itertuples():
    doc = rec.full_text
    doc = preprocessing_pipeline(doc)
    doc = doc.split("\n")
    for i, sentence in enumerate(doc):
        sentence_train_df.loc[cnt, "text_id"] = rec.text_id
        sentence_train_df.loc[cnt, "sentence_id"] = i
        sentence_train_df.loc[cnt, "sentence"] = sentence
        for colname in target_colnames:
            sentence_train_df.loc[cnt, colname] = getattr(rec, colname)
        cnt += 1
```

[ ]:

```
sentence_train_df.head()
```

[ ]:

```
        text_id  sentence_id  \
0  0016926B079C          0.0
1  0016926B079C          1.0
2  0016926B079C          2.0
3  0016926B079C          3.0
4  0022683E9EA5          0.0

                                            sentence  cohesion  syntax  \
0  I think that students would benefit from learn…       3.5     3.5
1  The hardest part of school is getting ready. y…       3.5     3.5
2  most students usually take showers before scho…       3.5     3.5
3  when your home your comfortable and you pay at…       3.5     3.5
4  When a problem is a change you have to let it …       2.5     2.5

   vocabulary  phraseology  grammar  conventions
0         3.0          3.0      4.0          3.0
1         3.0          3.0      4.0          3.0
2         3.0          3.0      4.0          3.0
3         3.0          3.0      4.0          3.0
4         3.0          2.0      2.0          2.5
```

[ ]:

```
def tokenize(x):
    return len(tokenizer.encode(x))

sentence_train_df["subword_count"] = sentence_train_df["sentence"].apply(lambda␣
 ↪x: tokenize(x))
```

```
Token indices sequence length is longer than the specified maximum sequence
length for this model (1264 > 512). Running this sequence through the model will
result in indexing errors
```

[ ]:

```
sentence_train_df.head()
```

[ ]:

```
        text_id  sentence_id  \
0  0016926B079C          0.0
```

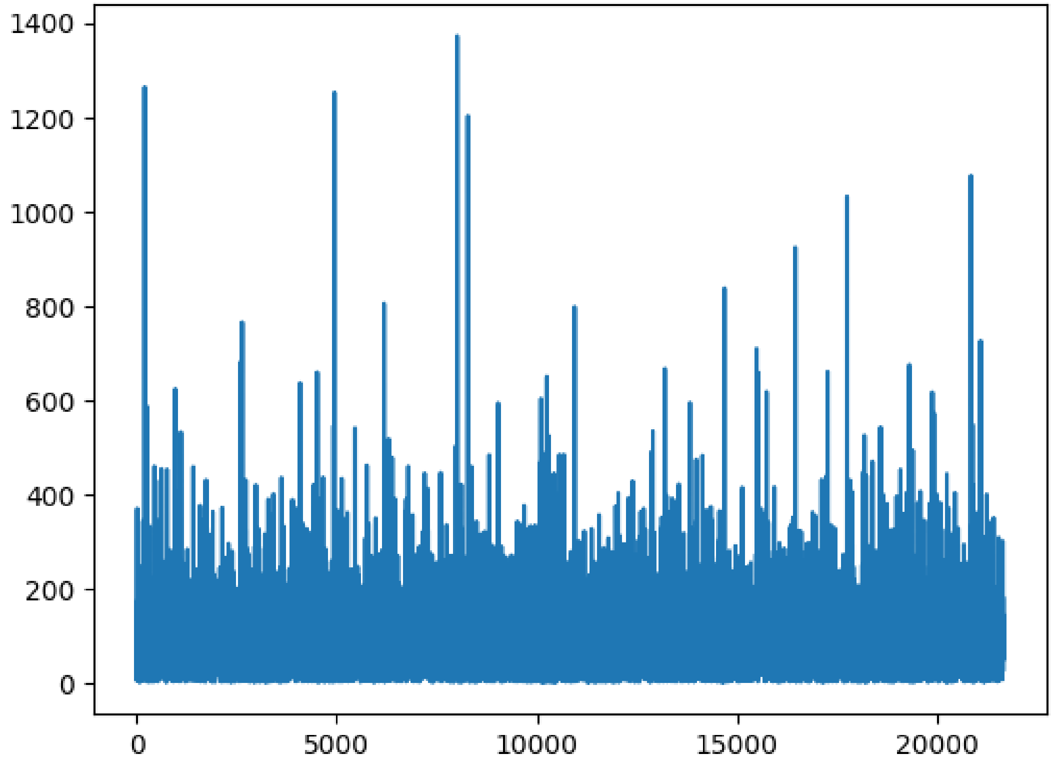
```
1  0016926B079C          1.0
2  0016926B079C          2.0
3  0016926B079C          3.0
4  0022683E9EA5          0.0

                                          sentence  cohesion  syntax  \
0  I think that students would benefit from learn…       3.5     3.5
1  The hardest part of school is getting ready. y…       3.5     3.5
2  most students usually take showers before scho…       3.5     3.5
3  when your home your comfortable and you pay at…       3.5     3.5
4  When a problem is a change you have to let it …       2.5     2.5

   vocabulary  phraseology  grammar  conventions  subword_count
0         3.0          3.0      4.0          3.0             58
1         3.0          3.0      4.0          3.0             89
2         3.0          3.0      4.0          3.0             85
3         3.0          3.0      4.0          3.0             67
4         3.0          2.0      2.0          2.5             59
```

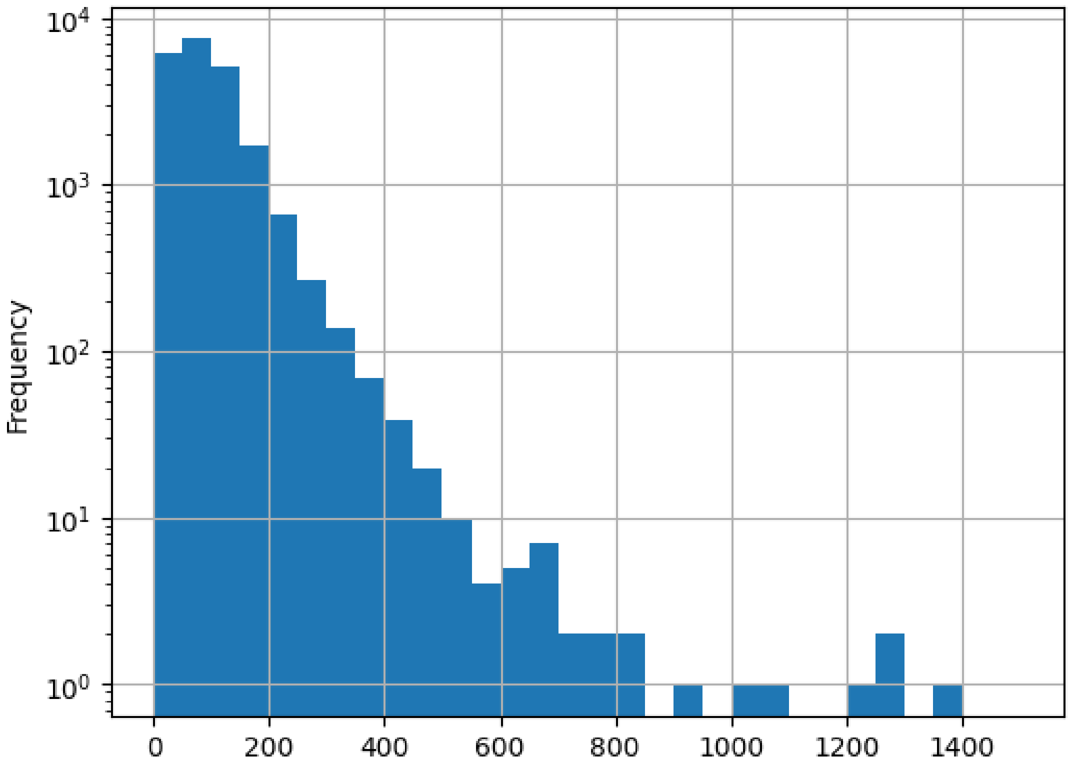
[ ]:
```
sentence_train_df["subword_count"].plot()
```

[ ]:
```
<AxesSubplot:>
```



[ ]:

```
n = 50
fig, axs = plt.subplots()
sentence_train_df["subword_count"].plot.hist(bins=np.arange(0, 1500 + n, n),␣
  ↪ax=axs)
axs.set_yscale("log")
axs.grid(True)
```



[ ]:

```
sentence_train_df["subword_count"].describe()
```

[ ]:

```
count    21661.000000
mean        90.490651
std         69.618733
min          3.000000
25%         44.000000
50%         81.000000
75%        119.000000
max       1372.000000
Name: subword_count, dtype: float64
```

[ ]:

```
sentence_train_df["subword_count"].quantile(.9)
```

[ ]:

```
166.0
```

[ ]:
```
sentence_train_df["subword_count"].quantile(.95)
```

[ ]:
```
207.0
```

[ ]:
```
sentence_train_df["subword_count"].quantile(.99)
```

[ ]:
```
328.0
```

[ ]:
```
sentence_train_df["subword_count"].quantile(.995)
```

[ ]:
```
387.40000000000146
```

[ ]:
```
sentence_train_df["subword_count"].quantile(.999)
```

[ ]:
```
620.0400000000227
```

[ ]:
```
sentence_train_df["subword_count"].quantile(.9975)
```

[ ]:
```
459.8500000000022
```

[ ]:
```
sentence_train_df.sort_values("subword_count", ascending=False).head(25)
```

[ ]:
```
            text_id  sentence_id  \
8015   6D114CB7EBB3          0.0
204    0369AD4B5726          0.0
4942   424005E31A04          2.0
8279   71F12F9F01DE          1.0
20856  F9E1434B2151          0.0
17755  E0656353B8F2          0.0
16463  D3A6B04B54A9          0.0
14695  C2FBC10F36E0          0.0
6187   54304ACFE477          0.0
10947  92FFB3530304          1.0
2638   22E55C32A973          0.0
21096  FB6F9B24FD1B          0.0
15498  CB71E5F9AE1E          0.0
2582   21EFAE02832D          0.0
19317  EDD899D4131D          0.0
13205  AF98232BDB08          0.0
17276  DBE977177080          8.0
4513   3DC3405DF531          0.0
15543  CBF62E8FFA1B          0.0
10256  89114CEF532E          7.0
4085   37193A72B5D2          1.0
964    0CB1237268E3          2.0
15748  CDBB3ACAD2C8          0.0
19885  F28D4ACF03AA          0.0
```

```
10106  876DD6D022A9          2.0

                                                sentence  cohesion  syntax  \
8015   The Character of beyone my control it take inf…       3.0     2.5
204    I would agree with being honest at all times b…       2.0     3.0
4942   when you talking around attitude is how we act…       3.0     4.0
8279   the people needs to change to see the real thi…       2.0     1.0
20856  Throughout the course of our history,education…       3.0     4.0
17755  School districts shouldn't give the three year…       3.5     2.5
16463  Should school have a 3-4 week break in the sum…       3.5     3.5
14695  so just because i couldn open the exhibit wind…       2.5     2.0
6187   when we're born we come out of our mothers wom…       2.5     2.5
10947  if you Have a negative impression at first, an…       4.0     3.5
2638   Have you ever woke up one morning and thought …       2.5     3.5
21096  I think that students should commit to their c…       2.0     2.0
15498  Do you think failure is a good thing or bad th…       2.5     3.0
2582   I believe students should design summer projec…       2.0     2.0
19317  Generic_Name thinks the school should control …       2.5     3.0
13205  Some schools use cell phones in the classroom …       2.5     2.5
17276  This lesson is very important because many peo…       2.5     2.5
4513   all tho some people say online classes is good…       2.0     2.5
15543  In my opinion, I think that true self-esteem c…       3.0     3.5
10256  Generic_Name 's father was a soldier .Generic_…       2.5     2.5
4085   In my opinion I think different and I don't th…       3.5     2.5
964    but we should never quit trying. Everything ta…       2.5     3.0
15748  Honesty is a good policy because you are being…       2.5     3.5
19885  I have multiple talents and skills, but my fav…       2.5     3.5
10106  However other people think that self-esteem co…       4.0     3.0

       vocabulary  phraseology  grammar  conventions  subword_count
8015          3.0          2.5      2.0          2.0           1372
204           3.5          3.0      3.0          3.5           1264
4942          4.0          3.0      3.0          3.0           1252
8279          2.0          2.0      1.5          2.0           1202
20856         3.5          3.0      4.0          3.0           1076
17755         3.5          3.5      2.5          3.0           1032
16463         3.5          3.5      2.5          2.5            924
14695         3.0          3.0      2.5          2.0            837
6187          3.0          3.0      2.5          2.5            805
10947         3.0          3.0      3.0          3.0            798
2638          3.5          3.0      3.0          3.0            765
21096         3.0          2.0      2.0          2.5            726
15498         3.0          3.5      3.0          3.0            709
2582          3.0          3.0      2.0          3.0            680
19317         3.5          3.0      3.0          3.0            675
13205         3.0          2.5      2.5          2.5            667
17276         3.0          3.0      2.5          2.5            661
```

```
4513          3.0          2.5       2.0          2.0             659
15543         3.5          3.5       3.5          3.0             658
10256         2.5          3.0       2.0          2.5             650
4085          3.5          3.5       3.0          3.5             637
964           3.0          2.5       3.0          3.0             624
15748         3.0          3.0       3.0          3.5             618
19885         3.5          3.5       3.0          2.5             617
10106         3.5          2.5       3.0          2.5             603
```

```
[ ]: sentence_train_df.sort_values("subword_count", ascending=False).
      ↪iloc[0]["sentence"]
```

```
[ ]: "The Character of beyone my control it take influence to make it right. This
     prompt support that i am able to have my own character. The british naturalist
     is right because people make bad chooice sometime it dosen't mean they are bad.
     My character will show that how much i can do the right thing. Influence beyond
     control your mind but it mean you have to strong and kind to other. I control
     myself to do bad and good things i can only control me and my action that can be
     a issue. I make misstake and l can learn from my own mistack that are right.
     This is also supports how i act to other and be truth to myself. Sometime you
     don't have to be selfish but their many selfish human being that do the wrong
     things and they know what they doing even that were wrong. I tell myself that
     you are not perfact but doing the right things that isn't issue but the issue is
     when people take it wrong way and you can't do or said anythings. If i really
     things about that what is my character i'll able to said that i am not perfact
     but i do the right things and i care about other people feeling may hurt by me
     and action and influence that i can control myself to not make misstakes that
     i'll regrade later. i support that my character is influence because i control
     myself and the rest of is you body do the rest in science that are study
     knowledge about human body but it's nothing to do with this but your character
     tell you how to choose and be youself. Sometimes i am not proude of who i am but
     it's not about who you are it about what you do and and what you said. I can
     control myself to do the right thing and their many teenager that dosen't know
     how to be yourself because it takes lots of controling your character. Your
     character tell you to support your respond to other that you choose away you can
     control yourself. My character will show that how much i can control myself to
     be myself and not to be the person that other want me to be. I can't give up i
     have to be postivie to myself and other and not to things nagative went i have
     to thing postivie to myself. I do agree with this quote because my character
     will be what i can just be myself and choose the right thing to make it right.
     My character makes me who i am now even went i wasn't my self. My pass what make
     me more strong and to be brave to control my self to choose make it all the way.
     It wasn't easy i work really heard to be the person that i am now my jerueny the
     reasons that i know how to be able to controling my self. Life isn't easy to
     know everything you have to thougt your self to live in this world there many
     people even don't know what is to be able to love yourself and do the right
     thing. The Britesh naturalist and politician and John Lubbock wrote this quote
```

```
and i do believe that i can control what i said and i also control what are my
action are that not usless to me because it's importan to me by influence beyond
control take a position on this issue. For example i did losed control went i am
not thinking postivie and choosing to be nagative that's why people need to know
how to be able to control them self. Thinking postivie is't force it's come by
influences and the people you around with not always the you round with their
know how to control them sefl but it is not them that you can't depond on you
have to learn your own and not all people are bad to hanging out with. I believe
that one person can change the whold world just by their bravery and thinking
postivie to them self i may not be prefact but i know how to control myself just
by thinking postivie and think before i said or do anythigs. Some teen that are
don't know how to control them self but you really don't have to know how to
control your self they're many reasons influence that control youe self that you
don't have control over what you have is you control you i know went i frist
nocite that i have to make everything right but it is not what i have change it
is what i am. Character influence beyond and it takes many reasons to be able to
do the right thing i support that influence do control our character that we
have and it make it more hearder to believe that i know what is the right and
wrong thing as the time goes but my responds has to be the same i can't not
change what other people do or said but what i can do is just be myself and
think postivie that what makes it my character right. Character formed by
influence beyond my control i do have many ressons to support this prompt do i
choose my own character formed that is real by influences beyond control that i
have over myself is my action and my thoughts that able to do the right thing as
my character choose to be make it right. John Lubbock was right i do control my
character formed by influences beyond that my brain tells me to do. it is
positible to my character will be what me and myself choose to make it right, my
responds with this prompt that is i can control my character will be show as i
think postivie and do the right thing to improve my self that i can control my
to be make it thru all the way as i'm useing character formed by influences
beyond. I can control my actions and my thoughts that will not hurting other but
if they're someone is not nice to me i can just use my postivie thought and
postivie mind to control my nagaitve thoughts and use my postivie thoughts not
to be bad as their to me. I do also support this prompt i am controling myself
to be the person that i am and not to be the person that what other people want
you to be. My influences that control my action that i am able to do the right
thing and think postivie as same time my undestaning one anouther that support
that what other may feel. I choose to make it right because that is who i am and
it make me more stronger every time i am controling my action and that may not
be the problam as the respond stay the same ans iam supporting this because i
also do believe in myself that i do really can think postivie and do the right
things by influence beyone my control."
```

[ ]:

```
```

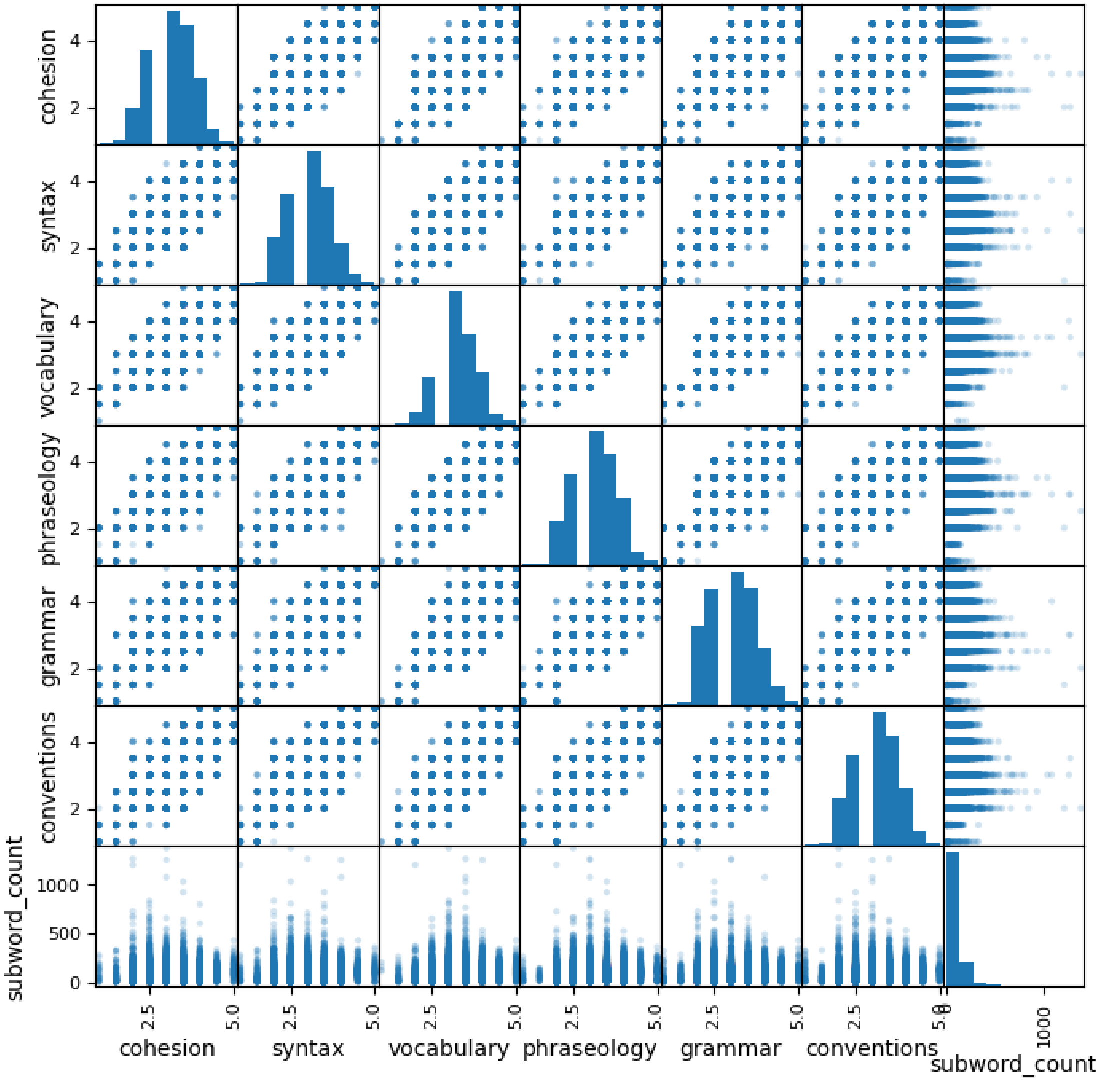
[ ]:

```
fig, axs = plt.subplots(figsize=(8,8))
```

```
pd.plotting.scatter_matrix(sentence_train_df[target_colnames +␣
 ↪["subword_count"]], alpha=0.2, ax=axs)
```

```
/home/miyakawa/workspace/kaggle/feedback-prize-english-language-
learning/.venv/lib/python3.7/site-packages/ipykernel_launcher.py:2: UserWarning:
To output multiple subplots, the figure containing the passed axes is being
cleared
```

```
[ ]: array([[<AxesSubplot:xlabel='cohesion', ylabel='cohesion'>,
             <AxesSubplot:xlabel='syntax', ylabel='cohesion'>,
             <AxesSubplot:xlabel='vocabulary', ylabel='cohesion'>,
             <AxesSubplot:xlabel='phraseology', ylabel='cohesion'>,
             <AxesSubplot:xlabel='grammar', ylabel='cohesion'>,
             <AxesSubplot:xlabel='conventions', ylabel='cohesion'>,
             <AxesSubplot:xlabel='subword_count', ylabel='cohesion'>],
            [<AxesSubplot:xlabel='cohesion', ylabel='syntax'>,
             <AxesSubplot:xlabel='syntax', ylabel='syntax'>,
             <AxesSubplot:xlabel='vocabulary', ylabel='syntax'>,
             <AxesSubplot:xlabel='phraseology', ylabel='syntax'>,
             <AxesSubplot:xlabel='grammar', ylabel='syntax'>,
             <AxesSubplot:xlabel='conventions', ylabel='syntax'>,
             <AxesSubplot:xlabel='subword_count', ylabel='syntax'>],
            [<AxesSubplot:xlabel='cohesion', ylabel='vocabulary'>,
             <AxesSubplot:xlabel='syntax', ylabel='vocabulary'>,
             <AxesSubplot:xlabel='vocabulary', ylabel='vocabulary'>,
             <AxesSubplot:xlabel='phraseology', ylabel='vocabulary'>,
             <AxesSubplot:xlabel='grammar', ylabel='vocabulary'>,
             <AxesSubplot:xlabel='conventions', ylabel='vocabulary'>,
             <AxesSubplot:xlabel='subword_count', ylabel='vocabulary'>],
            [<AxesSubplot:xlabel='cohesion', ylabel='phraseology'>,
             <AxesSubplot:xlabel='syntax', ylabel='phraseology'>,
             <AxesSubplot:xlabel='vocabulary', ylabel='phraseology'>,
             <AxesSubplot:xlabel='phraseology', ylabel='phraseology'>,
             <AxesSubplot:xlabel='grammar', ylabel='phraseology'>,
             <AxesSubplot:xlabel='conventions', ylabel='phraseology'>,
             <AxesSubplot:xlabel='subword_count', ylabel='phraseology'>],
            [<AxesSubplot:xlabel='cohesion', ylabel='grammar'>,
             <AxesSubplot:xlabel='syntax', ylabel='grammar'>,
             <AxesSubplot:xlabel='vocabulary', ylabel='grammar'>,
             <AxesSubplot:xlabel='phraseology', ylabel='grammar'>,
             <AxesSubplot:xlabel='grammar', ylabel='grammar'>,
             <AxesSubplot:xlabel='conventions', ylabel='grammar'>,
             <AxesSubplot:xlabel='subword_count', ylabel='grammar'>],
            [<AxesSubplot:xlabel='cohesion', ylabel='conventions'>,
             <AxesSubplot:xlabel='syntax', ylabel='conventions'>,
             <AxesSubplot:xlabel='vocabulary', ylabel='conventions'>,
             <AxesSubplot:xlabel='phraseology', ylabel='conventions'>,
             <AxesSubplot:xlabel='grammar', ylabel='conventions'>,
             <AxesSubplot:xlabel='conventions', ylabel='conventions'>,
             <AxesSubplot:xlabel='subword_count', ylabel='conventions'>],
            [<AxesSubplot:xlabel='cohesion', ylabel='subword_count'>,
             <AxesSubplot:xlabel='syntax', ylabel='subword_count'>,
             <AxesSubplot:xlabel='vocabulary', ylabel='subword_count'>,
             <AxesSubplot:xlabel='phraseology', ylabel='subword_count'>,
             <AxesSubplot:xlabel='grammar', ylabel='subword_count'>,
             <AxesSubplot:xlabel='conventions', ylabel='subword_count'>,
             <AxesSubplot:xlabel='subword_count', ylabel='subword_count'>]],
           dtype=object)
```



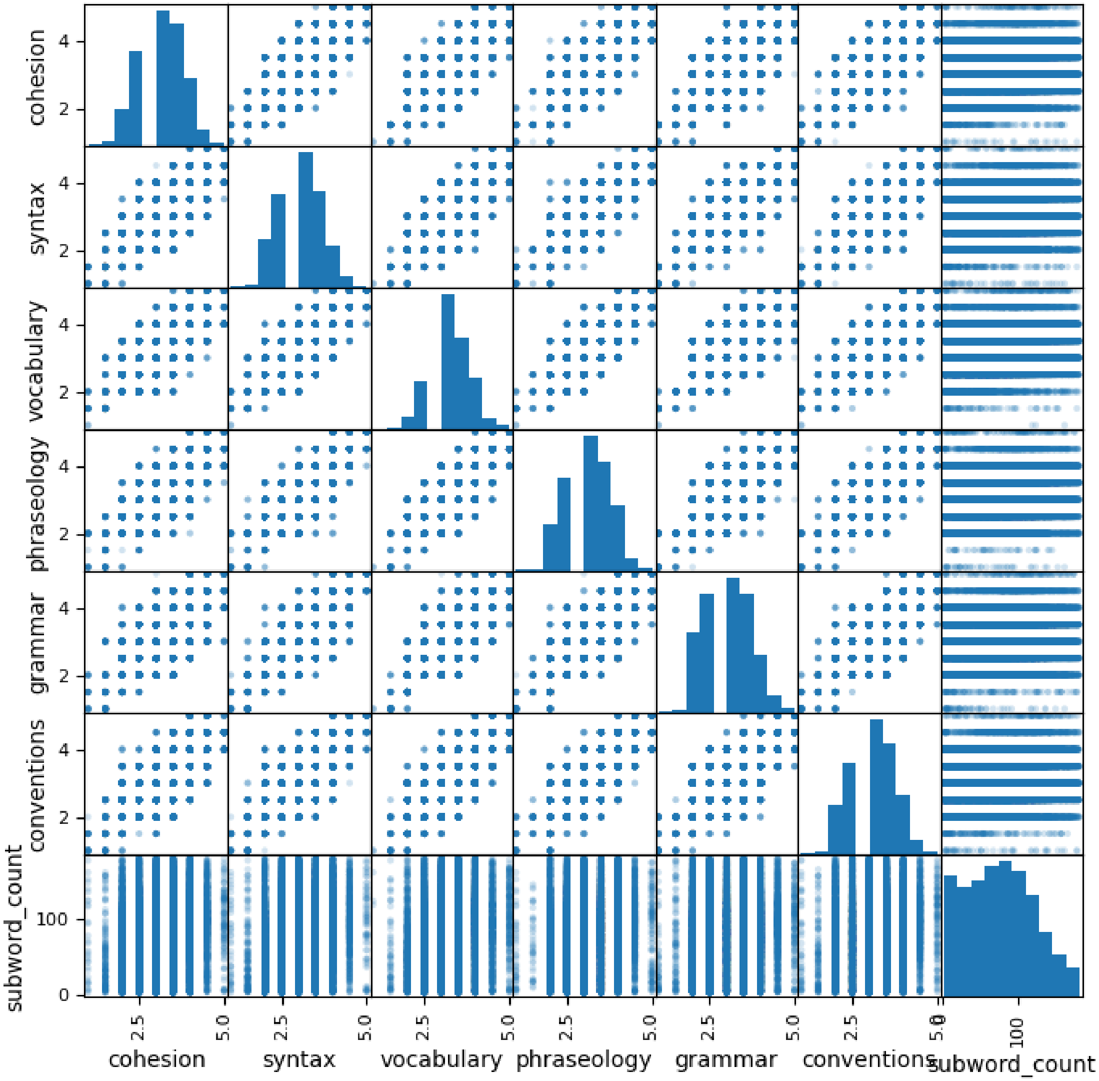
[ ]:

```
fig, axs = plt.subplots(figsize=(8,8))
pd.plotting.scatter_matrix(sentence_train_df.sort_values("subword_count").
 ↪head(20000)[target_colnames + ["subword_count"]], alpha=0.2, ax=axs)
```

```
/home/miyakawa/workspace/kaggle/feedback-prize-english-language-
learning/.venv/lib/python3.7/site-packages/ipykernel_launcher.py:2: UserWarning:
To output multiple subplots, the figure containing the passed axes is being
cleared
```

[ ]:

```
array([[<AxesSubplot:xlabel='cohesion', ylabel='cohesion'>,
        <AxesSubplot:xlabel='syntax', ylabel='cohesion'>,
        <AxesSubplot:xlabel='vocabulary', ylabel='cohesion'>,
        <AxesSubplot:xlabel='phraseology', ylabel='cohesion'>,
        <AxesSubplot:xlabel='grammar', ylabel='cohesion'>,
        <AxesSubplot:xlabel='conventions', ylabel='cohesion'>,
        <AxesSubplot:xlabel='subword_count', ylabel='cohesion'>],
       [<AxesSubplot:xlabel='cohesion', ylabel='syntax'>,
        <AxesSubplot:xlabel='syntax', ylabel='syntax'>,
        <AxesSubplot:xlabel='vocabulary', ylabel='syntax'>,
        <AxesSubplot:xlabel='phraseology', ylabel='syntax'>,
        <AxesSubplot:xlabel='grammar', ylabel='syntax'>,
        <AxesSubplot:xlabel='conventions', ylabel='syntax'>,
        <AxesSubplot:xlabel='subword_count', ylabel='syntax'>],
       [<AxesSubplot:xlabel='cohesion', ylabel='vocabulary'>,
        <AxesSubplot:xlabel='syntax', ylabel='vocabulary'>,
        <AxesSubplot:xlabel='vocabulary', ylabel='vocabulary'>,
        <AxesSubplot:xlabel='phraseology', ylabel='vocabulary'>,
        <AxesSubplot:xlabel='grammar', ylabel='vocabulary'>,
        <AxesSubplot:xlabel='conventions', ylabel='vocabulary'>,
        <AxesSubplot:xlabel='subword_count', ylabel='vocabulary'>],
       [<AxesSubplot:xlabel='cohesion', ylabel='phraseology'>,
        <AxesSubplot:xlabel='syntax', ylabel='phraseology'>,
        <AxesSubplot:xlabel='vocabulary', ylabel='phraseology'>,
        <AxesSubplot:xlabel='phraseology', ylabel='phraseology'>,
        <AxesSubplot:xlabel='grammar', ylabel='phraseology'>,
        <AxesSubplot:xlabel='conventions', ylabel='phraseology'>,
        <AxesSubplot:xlabel='subword_count', ylabel='phraseology'>],
       [<AxesSubplot:xlabel='cohesion', ylabel='grammar'>,
        <AxesSubplot:xlabel='syntax', ylabel='grammar'>,
        <AxesSubplot:xlabel='vocabulary', ylabel='grammar'>,
        <AxesSubplot:xlabel='phraseology', ylabel='grammar'>,
        <AxesSubplot:xlabel='grammar', ylabel='grammar'>,
        <AxesSubplot:xlabel='conventions', ylabel='grammar'>,
        <AxesSubplot:xlabel='subword_count', ylabel='grammar'>],
       [<AxesSubplot:xlabel='cohesion', ylabel='conventions'>,
        <AxesSubplot:xlabel='syntax', ylabel='conventions'>,
  <AxesSubplot:xlabel='vocabulary', ylabel='conventions'>,
  <AxesSubplot:xlabel='phraseology', ylabel='conventions'>,
  <AxesSubplot:xlabel='grammar', ylabel='conventions'>,
  <AxesSubplot:xlabel='conventions', ylabel='conventions'>,
  <AxesSubplot:xlabel='subword_count', ylabel='conventions'>],
 [<AxesSubplot:xlabel='cohesion', ylabel='subword_count'>,
  <AxesSubplot:xlabel='syntax', ylabel='subword_count'>,
  <AxesSubplot:xlabel='vocabulary', ylabel='subword_count'>,
  <AxesSubplot:xlabel='phraseology', ylabel='subword_count'>,
  <AxesSubplot:xlabel='grammar', ylabel='subword_count'>,
  <AxesSubplot:xlabel='conventions', ylabel='subword_count'>,
  <AxesSubplot:xlabel='subword_count', ylabel='subword_count'>]],
dtype=object)
```



```
[ ]:
```