```
Token
Stem
Lemma
Some stats
No of word count
Some kind insights
```

```
para="""
Reviewer 2 measured target lesion and made more reliable assessment.
reviewer 2 did not define any target lesions. reviewer 1 did - cervical lymph node left,
which  responded to therapy, in TP 2 less than 30% (SD)  in TP3 more than 30% (PR).
At TP 4 there are new enlarged celiac and retroperitoneal lymph nodes that further enlarge
at TP 5 - consistent with PD.
new hepatic lesion can indeed be detected in TP 3 (only detectable in arterial phase -
difficult to see)
CR with resolution of visible esophageal lesion
not easy to decide. lesion is gone in ct at TP2 (condition for CR). on barium swallow
esophagus is slightly rigid (slightly rigid is allowed for CR), no mucosal defects and
barium passes smoothly (also condition for CR). what is missing in the barium swallow
report is the ratio of upper esophagus part to narrow part. when measured by oneself the
ratio in TP2 and 3 is more than 3:2 - what prevents CR. Hence right decision is NN
At TP 6 right para tracheal lymph node that had decreased in size enlarges again, with
further growth at TP 7 - consistent with PD.
I agree with reviewer 1 assessment of presence of tumor burden at TP 4 and also new
equivocal lesion identified by reviewer1.
Reviewer 2 selection of lesion and measurement appears accurate.
PD not present at TP4...the retrotracheal LN marked as enlarging by reviewer 1 is smaller
than BL
There is increased esophageal wall thickening at site of primary esophageal lesion on TP3,
c/w PD
rev1
there are multiple new lung lesions at TP 2, therefore PD starting at TP 2 is correct
not the same patient
Partial response is the correct evaluation - decreasing size of nodal lesions.
lung lesion measured by reviewer 2 is not necessarily a metastasis, therefore I would
prefer review by reviewer 1.
Reviewer 2 measured to large and slightly different position to baseline
correct is PD
bone lesion is growing but there is no real soft tissue component visible in the bone
lesion.
Rev 2 market a bone lesion without a soft tissue component which should not be done
according to study guidelines
PD at TP2 - significant increasing size of lesions
Retroperitoneal LNs are measurable at BL and can be chosen as a TL, therefore I would
prefer the review by reviewer 2.
Reviewer measured the iliac vein in addition toi the lesion. The measurement of Reviewer 1
is exact
reviewer 1 is correct, there is a significant increase of the paraaortic LN at TP 4, and
there seem to be new LN mets, PD at TP 4 is correct
there is a significant increase of the LNs iliac right, PD at TP 2 is correct
more non specific fluid but not sure that pleural effusion is truly tumor related
initial 130 mm TP4: 157 cm
there is a significant progression of multiple retroperitoneal LNs, clear PD at TP 2
there are no new liver lesions at TP 2. the hypodense area in liver segment 4 is due to
fatty liver tissue, not a real lesion
There is progression as marked
new metastatic LNs since TP 3, review by reviewer 2 is preferable
```

```
progressive liver lesions already at tp 3
There is BL disease and the PD
lesion in the left adrenal gland is most likely benign
only non target lesion groups visible
Lesion in liver is calcified and seems NOT to be a NTL, therefore ND at BL and at the
following TPs is correct
images are evaluable
it is not for sure that there are really new lesions in the liver at TP 2. At TP there is
a good arterial phase which is missing at BL, at BL lesions might just not have been seen
due to this missing arterial phase
I would not have marked the prostate but the pleural/soft tissue component of bony met is
real
Incredibly subtle findings but present
no lesion out of the prostate
there is a target lesion like reviewer 2 described
PD at TP 2 due to increasing size of lesions
new lung lesions since TP 2, the decision is based on whether these are seen as new lung
mets. As they persist at TP 3 it is more likely that they are new mets an not of
inflammatory origin.
no clear disease according to RECST 1.1 visible
I don't primarily measure a intraprostatic lesions.
there is disease at Baseline - and a stable disease situation at TP 2.
lesions are progressive
At BL there is no clearly measurable lesion in the prostate, therefore I would prefer the
evaluation of reviewer 1.
no clear disease visible
Fluid is no unequivocal lesion / therefore no progression
difference is due to different choice of TLs at BL, I think that the choice of reviewer 2
is more convincing and there is clear progression at TP 2
Has seen Prostate lesion earlier than Reviewer 2
Progression of T and NT
new nodal lesions iliacal left side and soft tissue component of bone lesionis
Still PR at TP 3 is correct
Growth of retroperitoneal NTL LNs at TP 2 compared to BL is significant, therefore PD at
TP 2 is correct.
there is a pelvic  lesion like reviewer 2 stated
PD at TP 4 is correct
new fluid is no unequivocal lesion
Nothing that I would take as target.  measured in long axis
PD at TP 2 is the correct evaluation due to increasing liver lesions
no disease.  seminal vesicles
increasing size of bladder lesion
After carefully reviewing the imaging material I do not see a significant growthy of the
liver lesions at TP 3 and 4. Review by reviewer 2 should be preferred.
nodal lesions are morphological suspect
increasing number and size of liver lesions
I think these are not unequivocal increase.  I think there is likely an increase but I
would have put NN.  This is close and I understand the call.
TL% at tp2 do not warrant PD.
R2 provides appropriate response, unequivocal PD.
new adenopathy as described
TL chosen are more easily reproduced over time
Again TL are slightly better and more reproducible
TL are slightly better than R1
R1 measurements are more similar over time
reliable measures over all tp
No definite peritoneal disease
There is  a new lucency in L1 at timepoint two which reviewer 1 marked as a new lesion and
called PD. But, I would consider it an equivocal lesion as this could represent a benign
compression fracture deformity. No other new lesions were seen. I would agree with partial
response.
```

```
reviewer 2's target lesions are uniformly slightly over measured.  Do not agree with
unequiv progression of disease
agree with PD at tp6
agree with pd at tp6
agree with pd at tp6
TL chosen by review 1 are slightly better than TL for reviewer 2.
the TL did increase over time compatible with Pd
TL are increasing overtime
"
PD at tp4 forward is correct "
the intramuscular metastases noted by reviewer 1 as a new lesion were present at BL
intramuscular metastasis was present at BL
intramuscular mets present at BL
intramuscular mets were present at BL
intramuscular mets present at BL
Agree with assessment of target lung and para-aortic nodal lesions and BOR of SD (based on
more consistent measurements by reader 1).
Agree with assessment of target lung and para-aortic nodal lesions and BOR of SD (based on
more consistent measurements by reader 1).
New lung lesions at TP 5, the bone lesions at TP 4 are equivocal
Sufficient evidence to declare PD at TP4, based on growth in target lesions.
agree with reviewer 1's more robust selection of measurable lesions
more robust selection of target lesions
continue to agree with reader 1 of PR as BOR
new adenopathy at tp2
Unequivocal PD based on progression in several mediastinal nodes.
PD at TP2 and 3 is reasonable as several nodes demonstrate 50+% growth (particularly
paraesophageal and retrocrural nodes).
R2 has more optimal target lesion selection
R2 has more optimal target lesion selection
BL-TP9: R1 has more optimal target lesion measurements; disagree with PD at TP9, as per R2
BL-TP14: R1 has more optimal target lesion measurements; disagree with PD at TP9, as per
R2
R2 has more optimal detection of new lesions
R2 has more optimal detection of new lesions
BL-TP7: R2 has more optimal detection of NLs
BL-TP10: R2 has more optimal detection of NLs
R1 provides appropriate response. No unequivocal evidence of earlier PD.
reviewer 1's measurements are slightly more accurate at baseline
Do not agree with unequiv PD due to tiny nodules at tp9 (may be inflammatory)
overall agree with reviewer 1
overall agree with lesions measurements and analysis of reader 1
"overall agree with reader 1
"
agree with new adrenal met at tp3/PD
Agree with PD at tp3
R1 bases the decision to steer away from PD because the enlarging nodes which required it
later shrank.  Of course there is the possibility that they were reactive (as he/she
claims), but the chest was full of metastatic nodes at baseline, and it's a stretch to
claim that the enlarged mediastinal nodes were due to two separate processes.  We've all
seen tumor-bearing nodes (or non-nodal lesions, or new lesions) appear and then shrink
without thinking that the shrinking precluded tumor.  They COULD have been reactive, of
course - we can never know for sure - but they were most likely malignant.
Same rationale as that in Reason for Selection for Review Period 1.
Same reasons as for selecting Reviewer 2 at Review Period 1 and Review Period 2.
the RP adenopathy has increased greater than 20% at TP2
Adenopathy has increased from baseline PD is correct BOR
Better assessment of non targets
Better assessment of non targets
Better assessment of non targets
Better assessment of non targets
Better assessment of non targets
Better assessment of non targets
Better assessment of non targets
the right adrenal lesion is new from baseline.  However, the R adrenal was present at TP3
should have been equivocal and then updated to PD at TP3.  BOR would be SD. at TP2
I  agree with reviewer 1's new lesions at timepoint 2
Reviewer 1's new lesions at timepoint 2 are real.
As strange as it may seem to find a new liver lesion at timepoint 2 when all other lesions
are shrinking, I think it's unequivocal, and therefore requires PD at timepoint 2 and
thereafter.
The reviewers differ with regard to when PD might have first occurred, and have revised
their initial assessments.  But there has been PD, and for overall assessments at global
radiology review once PD has been assessed the overall assessments must remain PD.
the lesion reviewer 2 marked as new at TP3 was not new and had been present at TP2
PD at TP4
PD at TP4
Impressive. I would not have measured the lesion at first - thinking it was
intraprostatic.  It is obviously growing.  Well done.
no convincing TLs, review by reviewer 2 should be preferred
significant growth of soft tissue component of bone lesion
new intrapulmonary lesions
PD at TP3 is correct
no clear disease visible
pleural effusion and ascites since TP 2, PD is correct
its SD in TP 2 and TP 3 - therefore reviewer 2 is more correct
Yes, there are new liver mets since TP 2, PD at TP 2
All likely benign
Measurements by reviewer 2 seems more precise, should be preferred
Clear progression starts at TP 4.
very small but still visible soft tissue component of bone lesion - therefore SD is the
preferred valuation
no clear disease visible
leisions and PD is real
it is correct that there are new metastatic mesenteric LNs starting at TP 4
stable disease at TP 2 is correct
no clear measurable disease at BL
agreed with reviewer 1
agreed with reviewer 2
rev2
I Agree With OQREV1 For RP1.
I Agree with OQ REV1 for RP2.
more accurate TL measurement/selection. Reviewer 2 selected ill-defined peritoneal caking
as a TL
more accurate TL measurement/selection
there isn't clearly PD based on increased peritoneal disease at TP11
agree with lesion selection and measurements of reviewer 2
agree with measurements and lesion selection of reviewer 1
in retrospect, agree that their are three lung lesions, instead of 2 lung lesions plus a
hilar node.
R2 provides appropriate response.
R2 provides appropriate response.
R2 provides appropriate response.
BL-TP17: R2 has more optimal NT assessment
there is PD at TP 6 - newly enlarged R hilar LN
there is a new lesion in the left lower lobe which looks similar to the other metastatic
disease
Increased lung NT
Agree with PD at tp5
Agree with consistency of measurements by reviewer 2 for solitary target lesion. SD is
most appropriate at TP4.
```

```
Agree with consistency of measurements by reviewer 2 for solitary target lesion. SD is
most appropriate at TP4.
Although SD is perhaps more appropriate early in the assessment, PR is clearly present
from TP5 on and BOR of PR is appropriate from TP4 and beyond.
Although SD is perhaps more appropriate early in the assessment, PR is clearly present
from TP5 on and BOR of PR is appropriate from TP4 and beyond. Bone lesion in question by
reviewer 2 is overmeasured at later TPs.
insufficient # of new lung nodules to call it unequiv progression
agree with PR
PD at TP2 is correct based on new brain mets at TP2 and accompanying new sites of disease
at TP3.
differences related to techniques of measurement
Agree with PD identified by Reviewer 2
agree with BOR at tp3
in retrospect, agree with BOR at tp2
PD appropriate at TP3.
Periaortic nodes have been growing ever since baseline.  I might not have called them PD
until timepoint 3, but it's clear that PD occurred much closer to timepoint 2 than
timepoint 7.
Agree with PD at timepoint 2
Agree with PD at tp2
agree with PD at tp7
left adrenal nodule present at baseline.  agree with pd at tp7
agree with PD at tp7
both reviews are excellent; R1 has slightly better measurements.
TL measurements are slightly better
More targets selected and therefore more objective
More targets selected and therefore more objective
More targets selected and therefore more objective
the pleural lesion does increase at TP3 from nadir
new bone lesion best seen on bone scan and increase in the pleural lesion from nadir .
PD is correct at TP3
given the totality of the case I would change my opinion from R2 to R1.  the lesion that
were TL are well chosen and show evidence of decrease over time.  The new lesions chosen
by R2 do appear but could represent fractures.
"""
```

In [9]:

```python
import pandas as pd
```

In [10]:

```
df=pd.read_csv("xyz2.csv",encoding='latin-1')
df
```

Out[10]:

	ï»¿Reviewer 2 measured target lesion and made more reliable assessment.
0	reviewer 2 did not define any target lesions. ...
1	At TP 4 there are new enlarged celiac and retr...
2	new hepatic lesion can indeed be detected in T...
3	CR with resolution of visible esophageal lesion
4	not easy to decide. lesion is gone in ct at TP...
...	...
195	More targets selected and therefore more objec...
196	the pleural lesion does increase at TP3 from n...
197	new bone lesion best seen on bone scan and inc...
198	PD is correct at TP3
199	given the totality of the case I would change ...

200 rows × 1 columns

In [11]:

```
df.head()
```

Out[11]:

	ï»¿Reviewer 2 measured target lesion and made more reliable assessment.
0	reviewer 2 did not define any target lesions. ...
1	At TP 4 there are new enlarged celiac and retr...
2	new hepatic lesion can indeed be detected in T...
3	CR with resolution of visible esophageal lesion
4	not easy to decide. lesion is gone in ct at TP...

In [12]:

```
df.head(7)
```

Out[12]:

	ï»¿Reviewer 2 measured target lesion and made more reliable assessment.
0	reviewer 2 did not define any target lesions. ...
1	At TP 4 there are new enlarged celiac and retr...
2	new hepatic lesion can indeed be detected in T...
3	CR with resolution of visible esophageal lesion
4	not easy to decide. lesion is gone in ct at TP...
5	At TP 6 right para tracheal lymph node that ha...
6	I agree with reviewer 1 assessment of presence...

In [14]:

```
df.rename(columns={'ï»¿Reviewer 2 measured target lesion and made more reliable assessment.
df.head(5)
```

Out[14]:

	text
0	reviewer 2 did not define any target lesions. ...
1	At TP 4 there are new enlarged celiac and retr...
2	new hepatic lesion can indeed be detected in T...
3	CR with resolution of visible esophageal lesion
4	not easy to decide. lesion is gone in ct at TP...

In [15]:

```
df.shape
```

Out[15]:

```
(200, 1)
```

In [16]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   text    200 non-null    object
dtypes: object(1)
memory usage: 1.7+ KB
```

In [17]:

```
# missing values
df.isnull().sum()
```

Out[17]:

```
text    0
dtype: int64
```

In [18]:

```
# check for duplicate values
df.duplicated().sum()
```

Out[18]:

```
19
```

In [19]:

```
# remove duplicates
df = df.drop_duplicates(keep='first')
```

In [21]:

```
df.duplicated().sum()
```

Out[21]:

```
0
```

In [22]:

```
df.shape
```

Out[22]:

```
(181, 1)
```

In [23]:

```
df['text'].value_counts()
```

Out[23]:

```
reviewer 2 did not define any target lesions. reviewer 1 did - cervical lymp
h node left, which  responded to therapy, in TP 2 less than 30% (SD)  in TP3
more than 30% (PR).
1
the RP adenopathy has increased greater than 20% at TP2
1
Better assessment of non targets
1
the right adrenal lesion is new from baseline.  However, the R adrenal was p
resent at TP3 should have been equivocal and then updated to PD at TP3.  BOR
would be SD. at TP2
1
I  agree with reviewer 1's new lesions at timepoint 2
1


..
increasing size of bladder lesion
1
After carefully reviewing the imaging material I do not see a significant gr
owthy of the liver lesions at TP 3 and 4. Review by reviewer 2 should be pre
ferred.
1
nodal lesions are morphological suspect
1
increasing number and size of liver lesions
1
given the totality of the case I would change my opinion from R2 to R1.  the
lesion that were TL are well chosen and show evidence of decrease over time.
The new lesions chosen by R2 do appear but could represent fractures.      1
Name: text, Length: 181, dtype: int64
```

In [24]:

```
df.describe()
```

Out[24]:

	text
count	181
unique	181
top	reviewer 2 did not define any target lesions. ...
freq	1

In [30]:

```
df['num_sentences'] = df['text'].apply(lambda x:len(nltk.sent_tokenize(x)))
df['num_characters'] = df['text'].apply(len)
df['num_words'] = df['text'].apply(lambda x:len(nltk.word_tokenize(x)))
```

```
C:\Users\DELL\AppData\Local\Temp/ipykernel_23880/2662147901.py:1: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pand
as.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-v
ersus-a-copy)
  df['num_sentences'] = df['text'].apply(lambda x:len(nltk.sent_tokenize
(x)))
C:\Users\DELL\AppData\Local\Temp/ipykernel_23880/2662147901.py:2: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pand
as.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-v
ersus-a-copy)
  df['num_characters'] = df['text'].apply(len)
C:\Users\DELL\AppData\Local\Temp/ipykernel_23880/2662147901.py:3: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pand
as.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-v
ersus-a-copy)
  df['num_words'] = df['text'].apply(lambda x:len(nltk.word_tokenize(x)))
```

In [31]:

```
df.head()
```

Out[31]:

	text	num_sentences	num_characters	num_words
0	reviewer 2 did not define any target lesions. ...	2	172	43
1	At TP 4 there are new enlarged celiac and retr...	1	120	23
2	new hepatic lesion can indeed be detected in T...	1	105	21
3	CR with resolution of visible esophageal lesion	1	47	7
4	not easy to decide. lesion is gone in ct at TP...	6	432	91

In [32]:

```
df[['num_characters','num_words','num_sentences']].describe()
```

Out[32]:

	num_characters	num_words	num_sentences
count	181.000000	181.000000	181.000000
mean	71.143646	13.723757	1.193370
std	68.565632	13.843608	0.650778
min	4.000000	1.000000	1.000000
25%	31.000000	6.000000	1.000000
50%	51.000000	9.000000	1.000000
75%	86.000000	17.000000	1.000000
max	600.000000	115.000000	6.000000

In [35]:

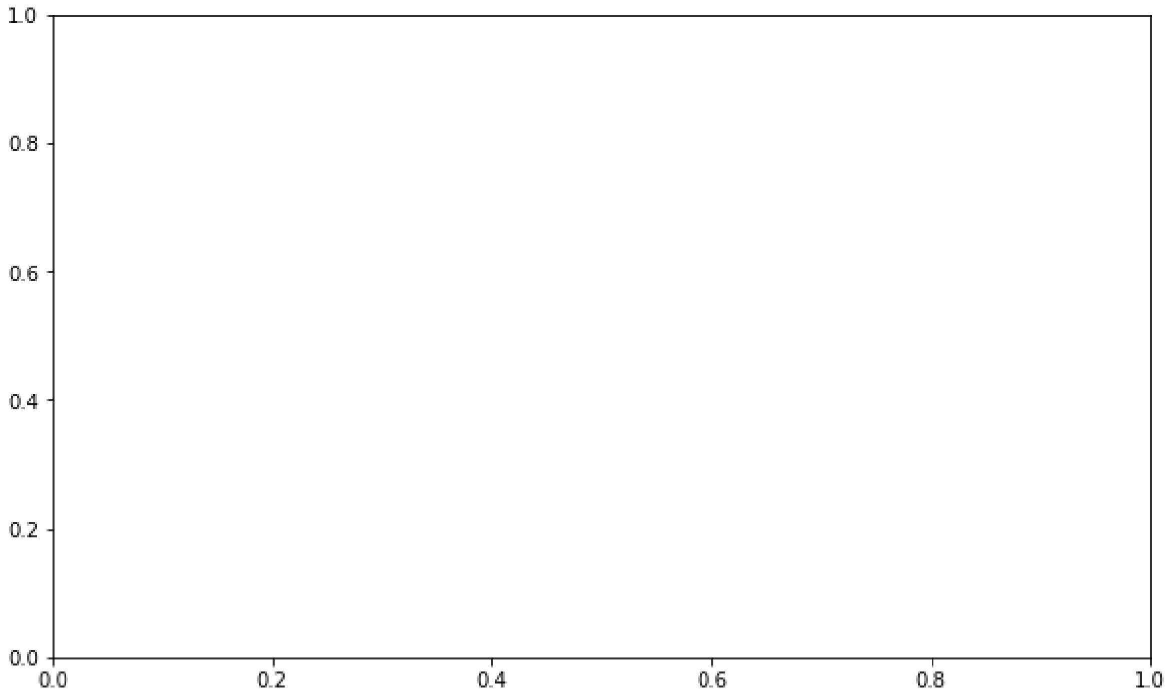
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [41]:

```
plt.figure(figsize=(10,6))
sns.histplot(df[df['text'] == 0]['num_characters'])
sns.histplot(df[df['text'] == 1]['num_characters'])
```

Out[41]:

```
<AxesSubplot:>
```



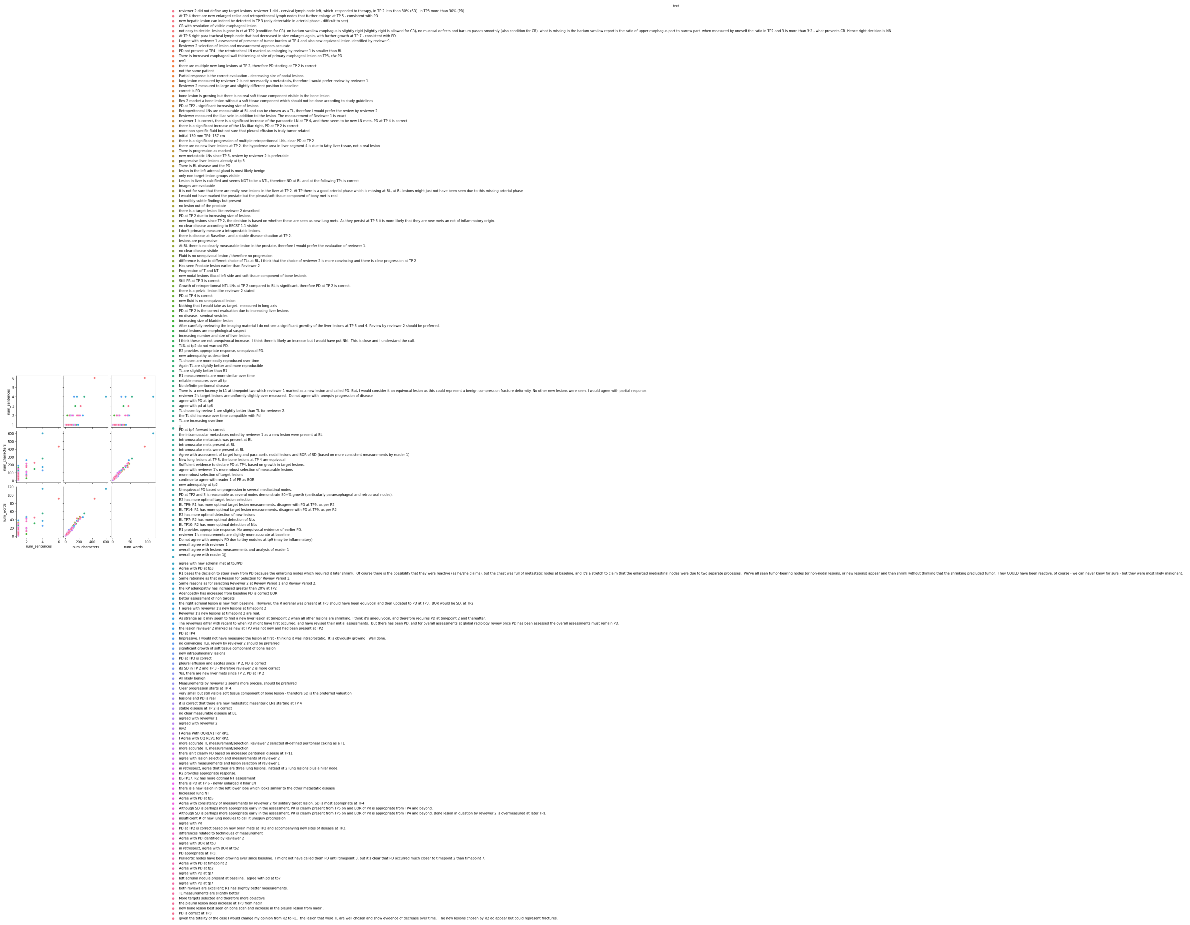
In [42]:

```
sns.pairplot(df,hue='text')
```

```
C:\Users\DELL\Anaconda3\lib\site-packages\matplotlib\backends\backend_agg.p
y:240: RuntimeWarning: Glyph 13 missing from current font.
  font.set_text(s, 0.0, flags=flags)
C:\Users\DELL\Anaconda3\lib\site-packages\matplotlib\backends\backend_agg.p
y:203: RuntimeWarning: Glyph 13 missing from current font.
  font.set_text(s, 0, flags=flags)
```

Out[42]:

```
<seaborn.axisgrid.PairGrid at 0x1f32b2be3a0>
```



In [27]:

```
import nltk
```

In [5]:

```
from nltk.tokenize import sent_tokenize

tokenized_sent=sent_tokenize(para)
print(tokenized_sent)
```

```
['\nReviewer 2 measured target lesion and made more reliable assessment.',
'reviewer 2 did not define any target lesions.', 'reviewer 1 did - cervica
l lymph node left, which  responded to therapy, in TP 2 less than 30% (SD)
in TP3 more than 30% (PR).', 'At TP 4 there are new enlarged celiac and re
troperitoneal lymph nodes that further enlarge at TP 5 - consistent with P
D.', 'new hepatic lesion can indeed be detected in TP 3 (only detectable i
n arterial phase - difficult to see) \nCR with resolution of visible esoph
ageal lesion\nnot easy to decide.', 'lesion is gone in ct at TP2 (conditio
n for CR).', 'on barium swallow esophagus is slightly rigid (slightly rigi
d is allowed for CR), no mucosal defects and barium passes smoothly (also
condition for CR).', 'what is missing in the barium swallow report is the
ratio of upper esophagus part to narrow part.', 'when measured by oneself
the ratio in TP2 and 3 is more than 3:2 - what prevents CR.', 'Hence right
decision is NN  \nAt TP 6 right para tracheal lymph node that had decrease
d in size enlarges again, with further growth at TP 7 - consistent with P
D.', 'I agree with reviewer 1 assessment of presence of tumor burden at TP
4 and also new equivocal lesion identified by reviewer1.', 'Reviewer 2 sel
ection of lesion and measurement appears accurate.', 'PD not present at TP
4...the retrotracheal LN marked as enlarging by reviewer 1 is smaller than
```

In [6]:

```
from nltk.tokenize import word_tokenize
tokenized_word=word_tokenize(para)
print(tokenized_word)
```

```
['Reviewer', '2', 'measured', 'target', 'lesion', 'and', 'made', 'more',
'reliable', 'assessment', '.', 'reviewer', '2', 'did', 'not', 'define', 'a
ny', 'target', 'lesions', '.', 'reviewer', '1', 'did', '-', 'cervical', 'l
ymph', 'node', 'left', ',', 'which', 'responded', 'to', 'therapy', ',', 'i
n', 'TP', '2', 'less', 'than', '30', '%', '(', 'SD', ')', 'in', 'TP3', 'mo
re', 'than', '30', '%', '(', 'PR', ')', '.', 'At', 'TP', '4', 'there', 'ar
e', 'new', 'enlarged', 'celiac', 'and', 'retroperitoneal', 'lymph', 'node
s', 'that', 'further', 'enlarge', 'at', 'TP', '5', '-', 'consistent', 'wit
h', 'PD', '.', 'new', 'hepatic', 'lesion', 'can', 'indeed', 'be', 'detecte
d', 'in', 'TP', '3', '(', 'only', 'detectable', 'in', 'arterial', 'phase',
'-', 'difficult', 'to', 'see', ')', 'CR', 'with', 'resolution', 'of', 'vis
ible', 'esophageal', 'lesion', 'not', 'easy', 'to', 'decide', '.', 'lesio
n', 'is', 'gone', 'in', 'ct', 'at', 'TP2', '(', 'condition', 'for', 'CR',
')', '.', 'on', 'barium', 'swallow', 'esophagus', 'is', 'slightly', 'rigi
d', '(', 'slightly', 'rigid', 'is', 'allowed', 'for', 'CR', ')', ',', 'n
o', 'mucosal', 'defects', 'and', 'barium', 'passes', 'smoothly', '(', 'als
o', 'condition', 'for', 'CR', ')', '.', 'what', 'is', 'missing', 'in', 'th
e', 'barium', 'swallow', 'report', 'is', 'the', 'ratio', 'of', 'upper', 'e
sophagus', 'part', 'to', 'narrow', 'part', '.', 'when', 'measured', 'by',
```

In [7]:

```
#words count
from nltk.probability import FreqDist
fdist=FreqDist(tokenized_word)
print(fdist)
```

```
<FreqDist with 589 samples and 2629 outcomes>
```

In [8]:

```
fdist.most_common(2)
```

Out[8]:

```
[('at', 97), ('.', 82)]
```

In [9]:

```
from nltk.corpus import stopwords
stop_words=set(stopwords.words("english"))
print(stop_words)
```

```
{'doesn', 'than', "shan't", 'which', 'couldn', 'yours', 'any', 'shan', 'hi
s', 'with', 'out', "should've", 'after', 'hers', "you've", 'same', 'too', 'w
on', 'below', 'own', 'don', 'into', "don't", 'what', 'been', 'had', 'most',
'not', 'mightn', 'and', 'before', 'me', 'these', 'll', 'm', 'y', 'again', "d
idn't", 'has', 'down', 'themselves', 'yourselves', 'further', "you'd", 'up',
'he', 'itself', 'that', 'be', 'being', 'have', 'i', "aren't", 'to', 'it', "n
eedn't", 'such', 'are', 'whom', 'hadn', 'few', "you're", 'doing', 'needn',
'where', 'by', 'against', "won't", "that'll", "mustn't", 'or', 'as', 've',
'wouldn', "shouldn't", 'ours', 'off', 'over', 'having', 'does', 'through',
'no', 'do', 's', 'if', 'them', 'is', "hasn't", "wouldn't", "doesn't", 'jus
t', 'the', 'during', 'because', 'on', 'now', 'in', 'mustn', 'here', 'how',
'd', 'haven', 'its', 'yourself', 'but', 'some', 'while', 'other', 'did', 'on
ly', 'him', 'once', 'this', 're', 'will', 'ourselves', "she's", 'they', 'wh
o', 'herself', 'didn', "isn't", "wasn't", 'why', 'at', 'a', 'hasn', 'until',
"haven't", 'ma', 'an', 'were', 'when', "it's", 'should', "you'll", 'so', 'sh
ouldn', 'their', 'there', 'o', 'those', 'all', 't', 'aren', "mightn't", 'was
n', "weren't", 'ain', 'under', "hadn't", 'of', 'you', 'we', 'each', 'our',
'myself', 'isn', 'more', 'theirs', 'weren', 'her', "couldn't", 'then', 'ca
n', 'himself', 'above', 'about', 'your', 'she', 'very', 'was', 'between', 'b
oth', 'nor', 'my', 'from', 'for', 'am'}
```

In [10]:

```
filtered_sent=[]
for w in tokenized_word:
    if w not in stop_words:
        filtered_sent.append(w)
print("ts:",tokenized_word)
print("fs:",filtered_sent)
```

```
ts: ['Reviewer', '2', 'measured', 'target', 'lesion', 'and', 'made', 'mor
e', 'reliable', 'assessment', '.', 'reviewer', '2', 'did', 'not', 'defin
e', 'any', 'target', 'lesions', '.', 'reviewer', '1', 'did', '-', 'cervica
l', 'lymph', 'node', 'left', ',', 'which', 'responded', 'to', 'therapy',
',', 'in', 'TP', '2', 'less', 'than', '30', '%', '(', 'SD', ')', 'in', 'TP
3', 'more', 'than', '30', '%', '(', 'PR', ')', '.', 'At', 'TP', '4', 'ther
e', 'are', 'new', 'enlarged', 'celiac', 'and', 'retroperitoneal', 'lymph',
'nodes', 'that', 'further', 'enlarge', 'at', 'TP', '5', '-', 'consistent',
'with', 'PD', '.', 'new', 'hepatic', 'lesion', 'can', 'indeed', 'be', 'det
ected', 'in', 'TP', '3', '(', 'only', 'detectable', 'in', 'arterial', 'pha
se', '-', 'difficult', 'to', 'see', ')', 'CR', 'with', 'resolution', 'of',
'visible', 'esophageal', 'lesion', 'not', 'easy', 'to', 'decide', '.', 'le
sion', 'is', 'gone', 'in', 'ct', 'at', 'TP2', '(', 'condition', 'for', 'C
R', ')', '.', 'on', 'barium', 'swallow', 'esophagus', 'is', 'slightly', 'r
igid', '(', 'slightly', 'rigid', 'is', 'allowed', 'for', 'CR', ')', ',',
'no', 'mucosal', 'defects', 'and', 'barium', 'passes', 'smoothly', '(', 'a
lso', 'condition', 'for', 'CR', ')', '.', 'what', 'is', 'missing', 'in',
'the', 'barium', 'swallow', 'report', 'is', 'the', 'ratio', 'of', 'upper',
'esophagus', 'part', 'to', 'narrow', 'part', '.', 'when', 'measured', 'b
```

In [1]:

```
from nltk.tokenize import sent_tokenize, word_tokenize
ps=PorterStemmer()
stemmed_words=[]
for w in filtered_sent:
    stemmed_words.append(ps.stem(para))
print(filtered_sent)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_20160/533033200.py in <module>
      1 from nltk.tokenize import sent_tokenize, word_tokenize
----> 2 ps=PorterStemmer()
      3 stemmed_words=[]
      4 for w in filtered_sent:
      5     stemmed_words.append(ps.stem(para))

NameError: name 'PorterStemmer' is not defined
```

In [2]:

```
from nltk.stem.wordnet import WordNetLemmatizer
ls=WordNetLemmatizer()
#from nltk.stem import PorterStemmer
#stem =PorterStemmer()
word="flying"
print("l_w;",ls.lemmatize(word,"g"))
```

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)
~\Anaconda3\lib\site-packages\nltk\corpus\util.py in __load(self)
     83                 try:
---> 84                     root = nltk.data.find(f"{self.subdir}/{zip_nam
e}")
     85                 except LookupError:

~\Anaconda3\lib\site-packages\nltk\data.py in find(resource_name, paths)
    582     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 583     raise LookupError(resource_not_found)
    584 

LookupError: 
**********************************************************************
  Resource wordnet not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('wordnet')
  
  For more information see: https://www.nltk.org/data.html (https://www.nlt
k.org/data.html)

  Attempted to load corpora/wordnet.zip/wordnet/

  Searched in:
    - 'C:\\Users\\DELL/nltk_data'
    - 'C:\\Users\\DELL\\Anaconda3\\nltk_data'
    - 'C:\\Users\\DELL\\Anaconda3\\share\\nltk_data'
    - 'C:\\Users\\DELL\\Anaconda3\\lib\\nltk_data'
    - 'C:\\Users\\DELL\\AppData\\Roaming\\nltk_data'
    - 'C:\\nltk_data'
    - 'D:\\nltk_data'
    - 'E:\\nltk_data'
**********************************************************************


During handling of the above exception, another exception occurred:

LookupError                               Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_20160/2970926820.py in <module>
      4 #stem =PorterStemmer()
      5 word="flying"
----> 6 print("l_w;",ls.lemmatize(word,"g"))

~\Anaconda3\lib\site-packages\nltk\stem\wordnet.py in lemmatize(self, word,
 pos)
     43         :return: The lemma of `word`, for the given `pos`.
     44         """
```

```
---> 45         lemmas = wn._morphy(word, pos)
     46         return min(lemmas, key=len) if lemmas else word
     47 

~\Anaconda3\lib\site-packages\nltk\corpus\util.py in __getattr__(self, attr)
    119             raise AttributeError("LazyCorpusLoader object has no att
ribute '__bases__'")
    120 
--> 121         self.__load()
    122         # This looks circular, but its not, since __load() changes o
ur
    123         # __class__ to something new:

~\Anaconda3\lib\site-packages\nltk\corpus\util.py in __load(self)
     84                     root = nltk.data.find(f"{self.subdir}/{zip_nam
e}")
     85                 except LookupError:
---> 86                     raise e
     87 
     88         # Load the corpus.

~\Anaconda3\lib\site-packages\nltk\corpus\util.py in __load(self)
     79         else:
     80             try:
---> 81                 root = nltk.data.find(f"{self.subdir}/{self.__name}"
)
     82             except LookupError as e:
     83                 try:

~\Anaconda3\lib\site-packages\nltk\data.py in find(resource_name, paths)
    581     sep = "*" * 70
    582     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 583     raise LookupError(resource_not_found)
    584 
    585 

LookupError: 
**********************************************************************
  Resource wordnet not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('wordnet')
  
  For more information see: https://www.nltk.org/data.html (https://www.nlt
k.org/data.html)

  Attempted to load corpora/wordnet

  Searched in:
    - 'C:\\Users\\DELL/nltk_data'
    - 'C:\\Users\\DELL\\Anaconda3\\nltk_data'
    - 'C:\\Users\\DELL\\Anaconda3\\share\\nltk_data'
    - 'C:\\Users\\DELL\\Anaconda3\\lib\\nltk_data'
    - 'C:\\Users\\DELL\\AppData\\Roaming\\nltk_data'
    - 'C:\\nltk_data'
    - 'D:\\nltk_data'
    - 'E:\\nltk_data'
**********************************************************************
```

In [13]:

```
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem.wordnet import WordNetLemmatizer
ls=WordNetLemmatizer()
lemm_words=[]
for w in filtered_sent:
    lemm_words.append(ls.lemmatize(para))

print(filtered_sent)
```

```
---------------------------------------------------------------------------
-
LookupError                               Traceback (most recent call las
t)
~\Anaconda3\lib\site-packages\nltk\corpus\util.py in __load(self)
     83                 try:
---> 84                     root = nltk.data.find(f"{self.subdir}/{zip_nam
e}")
     85                 except LookupError:

~\Anaconda3\lib\site-packages\nltk\data.py in find(resource_name, paths)
    582     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 583     raise LookupError(resource_not_found)
    584

LookupError:
**********************************************************************
  Resource wordnet not found.
  Please use the NLTK Downloader to obtain the resource:
```

In [14]:

```
pip install wordnet
```

```
Requirement already satisfied: wordnet in c:\users\dell\anaconda3\lib\site-p
ackages (0.0.1b2)Note: you may need to restart the kernel to use updated pac
kages.
Requirement already satisfied: colorama==0.3.9 in c:\users\dell\anaconda3\li
b\site-packages (from wordnet) (0.3.9)
```

In [15]:

```
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem.wordnet import WordNetLemmatizer
ls=WordNetLemmatizer()
lemm_words=[]
for w in filtered_sent:
    lemm_words.append(ls.lemmatize(para))

print(filtered_sent)
```

```
---------------------------------------------------------------------------
-
LookupError                               Traceback (most recent call las
t)
~\Anaconda3\lib\site-packages\nltk\corpus\util.py in __load(self)
     83                 try:
---> 84                     root = nltk.data.find(f"{self.subdir}/{zip_nam
e}")
     85                 except LookupError:

~\Anaconda3\lib\site-packages\nltk\data.py in find(resource_name, paths)
    582     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 583     raise LookupError(resource_not_found)
    584 

LookupError: 
**********************************************************************
  Resource wordnet not found.
  Please use the NLTK Downloader to obtain the resource:
```

In [ ]:

```

```