sts_test

December 8, 2021

How to clean the sts-test

sed -E -e 's/(Europe|StackExchange).*//' stsbenchmark/sts-test.csv > stsbenchmark/sts-test2.csv
Remove the " at in line 1118

[ ]:
```
import pandas as pd
import fasttext.util
import similaritymeasures
import string
```

[ ]:
```
sts_test = pd.read_csv('stsbenchmark/sts-test2.csv', sep="\t", header=None,␣
↪names=['genre', 'file', 'years', 'id', 'sim', 'sent1', 'sent2'])

sts_test
```

[ ]:
```
              genre       file     years    id  sim  \
0     main-captions     MSRvid  2012test    24  2.5
1     main-captions     MSRvid  2012test    33  3.6
2     main-captions     MSRvid  2012test    45  5.0
3     main-captions     MSRvid  2012test    63  4.2
4     main-captions     MSRvid  2012test    66  1.5
…               …          …         …     …  …
1374      main-news  headlines      2016  1354  0.0
1375      main-news  headlines      2016  1360  1.0
1376      main-news  headlines      2016  1368  1.0
1377      main-news  headlines      2016  1420  0.0
1378      main-news  headlines      2016  1432  0.0

                                                  sent1  \
0                            A girl is styling her hair.
1                  A group of men play soccer on the beach.
2           One woman is measuring another woman's ankle.
3                         A man is cutting up a cucumber.
4                                 A man is playing a harp.
…                                                      …
1374  Philippines, Canada pledge to further boost re…
1375  Israel bars Palestinians from Jerusalem's Old …
1376         How much do you know about Secret Service?
```

```
1377  Obama Struggles to Soothe Saudi Fears As Iran …
1378           South Korea declares end to MERS outbreak

                                                  sent2
0                            A girl is brushing her hair.
1       A group of boys are playing soccer on the beach.
2                 A woman measures another woman's ankle.
3                            A man is slicing a cucumber.
4                            A man is playing a keyboard.
…                                                     …
1374              Philippines saves 100 after ferry sinks
1375  Two-state solution between Palestinians, Israe…
1376  Lawmakers from both sides express outrage at S…
1377  Myanmar Struggles to Finalize Voter Lists for …
1378  North Korea Delegation Meets With South Korean…

[1379 rows x 7 columns]
```

[ ]:
```
# make sent1 and sent2 lower
sts_test['sent1'] = sts_test['sent1'].str.lower()
sts_test['sent2'] = sts_test['sent2'].str.lower()

# remove punctuation
sts_test['sent1'] = sts_test['sent1'].apply(lambda x: str(x).translate(str.
 ↪maketrans('', '', string.punctuation)))
sts_test['sent2'] = sts_test['sent2'].apply(lambda x: str(x).translate(str.
 ↪maketrans('', '', string.punctuation)))

sts_test
```

[ ]:
```
              genre       file     years    id  sim  \
0     main-captions     MSRvid  2012test    24  2.5
1     main-captions     MSRvid  2012test    33  3.6
2     main-captions     MSRvid  2012test    45  5.0
3     main-captions     MSRvid  2012test    63  4.2
4     main-captions     MSRvid  2012test    66  1.5
…               …          …         …     …   …
1374      main-news  headlines      2016  1354  0.0
1375      main-news  headlines      2016  1360  1.0
1376      main-news  headlines      2016  1368  1.0
1377      main-news  headlines      2016  1420  0.0
1378      main-news  headlines      2016  1432  0.0

                                                  sent1  \
0                            a girl is styling her hair
1                   a group of men play soccer on the beach
2             one woman is measuring another womans ankle
```

```
3                        a man is cutting up a cucumber
4                              a man is playing a harp
…                                                    …
1374  philippines canada pledge to further boost rel…
1375  israel bars palestinians from jerusalems old city
1376          how much do you know about secret service
1377  obama struggles to soothe saudi fears as iran …
1378          south korea declares end to mers outbreak

                                                  sent2
0                            a girl is brushing her hair
1       a group of boys are playing soccer on the beach
2                  a woman measures another womans ankle
3                            a man is slicing a cucumber
4                            a man is playing a keyboard
…                                                    …
1374            philippines saves 100 after ferry sinks
1375  twostate solution between palestinians israel …
1376  lawmakers from both sides express outrage at s…
1377  myanmar struggles to finalize voter lists for …
1378  north korea delegation meets with south korean…

[1379 rows x 7 columns]
```

[ ]:
```
# load fasttext model
ft = fasttext.load_model('cc.en.300.bin')
```

```
Warning : `load_model` does not return WordVectorModel or SupervisedModel any
more, but a `FastText` object which is very similar.
```

[ ]:
```
# create columns of word vectors
sts_test['sent1_wv'] = sts_test['sent1'].apply(lambda x: [ft.
 ↪get_word_vector(word) for word in x.split()])
sts_test['sent2_wv'] = sts_test['sent2'].apply(lambda x: [ft.
 ↪get_word_vector(word) for word in x.split()])
```

[ ]:
```
# find Frechet similarity
sts_test['frech_sim'] = sts_test.apply(lambda x: similaritymeasures.
 ↪frechet_dist(x.sent1_wv, x.sent2_wv), axis=1)
sts_test
```

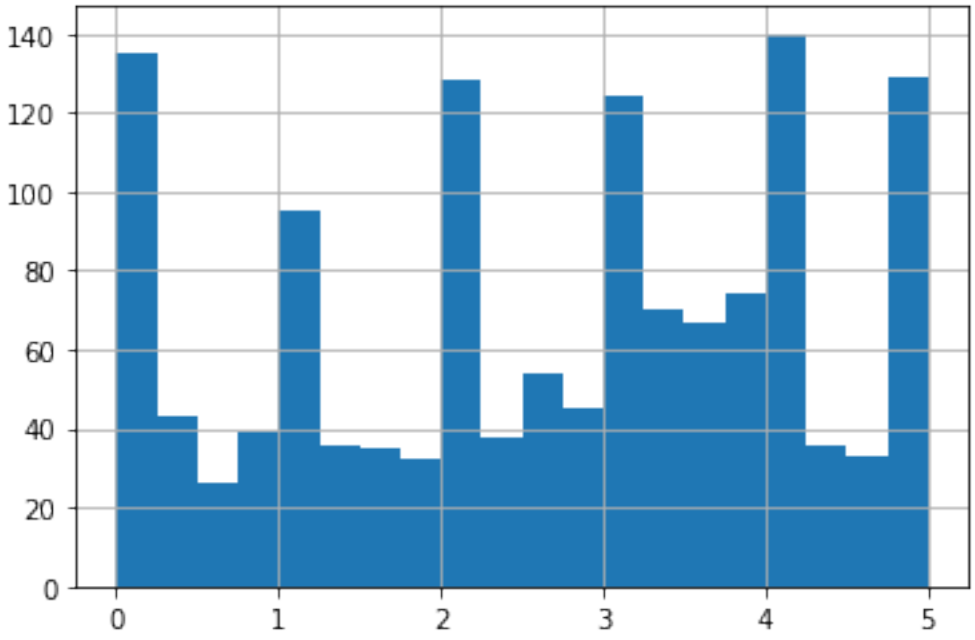
[ ]:
```
              genre       file     years    id  sim  \
0     main-captions     MSRvid  2012test    24  2.5
1     main-captions     MSRvid  2012test    33  3.6
2     main-captions     MSRvid  2012test    45  5.0
3     main-captions     MSRvid  2012test    63  4.2
4     main-captions     MSRvid  2012test    66  1.5
```

```
…              …         …        …    …  …
1374      main-news  headlines      2016  1354  0.0
1375      main-news  headlines      2016  1360  1.0
1376      main-news  headlines      2016  1368  1.0
1377      main-news  headlines      2016  1420  0.0
1378      main-news  headlines      2016  1432  0.0

                                                  sent1  \
0                            a girl is styling her hair
1                a group of men play soccer on the beach
2            one woman is measuring another womans ankle
3                        a man is cutting up a cucumber
4                               a man is playing a harp
…                                                     …
1374  philippines canada pledge to further boost rel…
1375  israel bars palestinians from jerusalems old city
1376          how much do you know about secret service
1377  obama struggles to soothe saudi fears as iran …
1378          south korea declares end to mers outbreak

                                                  sent2  \
0                            a girl is brushing her hair
1       a group of boys are playing soccer on the beach
2                  a woman measures another womans ankle
3                            a man is slicing a cucumber
4                            a man is playing a keyboard
…                                                     …
1374            philippines saves 100 after ferry sinks
1375  twostate solution between palestinians israel …
1376  lawmakers from both sides express outrage at s…
1377  myanmar struggles to finalize voter lists for …
1378  north korea delegation meets with south korean…

                                                sent1_wv  \
0     [[0.08764305, -0.49590126, -0.04985499, -0.093…
1     [[0.08764305, -0.49590126, -0.04985499, -0.093…
2     [[-0.0039191786, 0.03269286, -0.037494283, 0.2…
3     [[0.08764305, -0.49590126, -0.04985499, -0.093…
4     [[0.08764305, -0.49590126, -0.04985499, -0.093…
…                                                   …
1374  [[0.034213725, 0.052513517, -0.056771673, -0.0…
1375  [[-0.071692675, -0.052864145, -0.036141627, -0…
1376  [[-0.06619154, 0.0026334375, 0.113023736, -0.0…
1377  [[-0.12632462, -0.08979887, 0.012833006, 0.061…
1378  [[-0.077547535, 0.033660274, -0.006174694, 0.0…

                                                sent2_wv  frech_sim
```

```
0     [[0.08764305, -0.49590126, -0.04985499, -0.093…   0.958711
1     [[0.08764305, -0.49590126, -0.04985499, -0.093…   1.857392
2     [[0.08764305, -0.49590126, -0.04985499, -0.093…   2.968911
3     [[0.08764305, -0.49590126, -0.04985499, -0.093…   3.686892
4     [[0.08764305, -0.49590126, -0.04985499, -0.093…   1.533111
…                                                   …        …
1374  [[0.034213725, 0.052513517, -0.056771673, -0.0…   2.971272
1375  [[-0.01967115, 0.0038411804, 0.012567041, 0.09…   2.988335
1376  [[-0.08019948, 0.018645115, -0.02497993, -0.00…   3.376170
1377  [[-0.048643112, 0.02467984, -0.0019136805, -0…   2.819863
1378  [[-0.061924003, 0.039316084, 0.01719258, 0.090…   2.924450

[1379 rows x 10 columns]
```

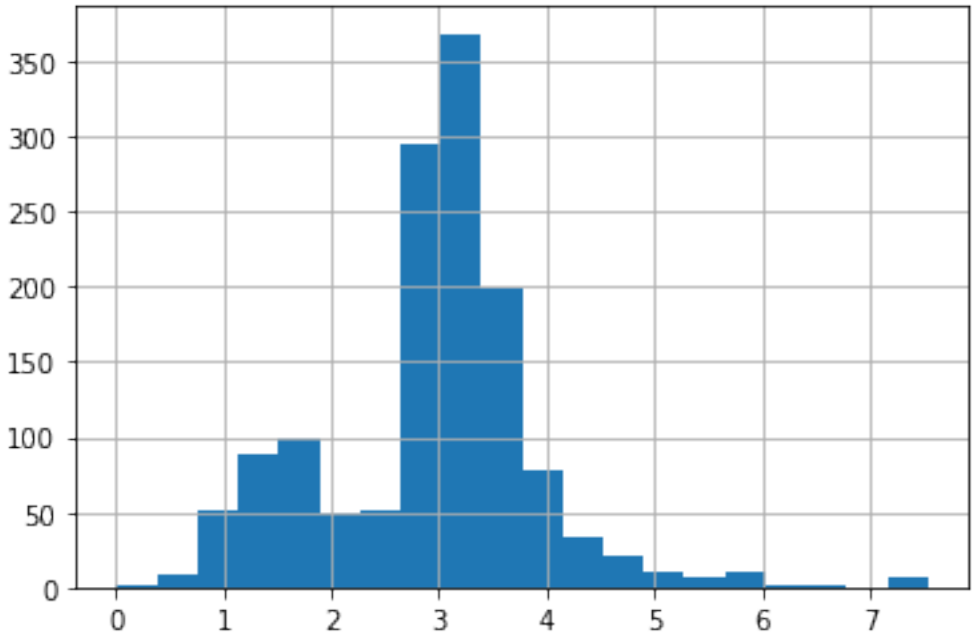
[ ]:
```
sts_test['sim'].hist(bins=20)
```

[ ]: <AxesSubplot:>



[ ]:
```
sts_test['frech_sim'].hist(bins=20)
```

[ ]: <AxesSubplot:>



[ ]:
```
print('sent1', sts_test[sts_test['frech_sim'] == sts_test['frech_sim'].
 ↪min()]['sent1'])
print('sent2', sts_test[sts_test['frech_sim'] == sts_test['frech_sim'].
 ↪min()]['sent2'])
```

```
sent1 623    a brown dog is jumping
Name: sent1, dtype: object
sent2 623    a brown dog is jumping
Name: sent2, dtype: object
```

[ ]:
```
print('sent1', sts_test[sts_test['frech_sim'] == sts_test['frech_sim'].
 ↪max()]['sent1'])
print('sent2', sts_test[sts_test['frech_sim'] == sts_test['frech_sim'].
 ↪max()]['sent2'])
```

```
sent1 1322    croatia begins countdown to historic eu entry
Name: sent1, dtype: object
sent2 1322    croatia countdowns to joining eu
Name: sent2, dtype: object
```

[ ]:
```
# False Positives: where the frechet similarity thinks they are similar but␣
 ↪they are not
sts_test[(sts_test['frech_sim'] <= 1) & (sts_test['sim'] < 2)][['sent1',␣
 ↪'sent2', 'frech_sim', 'sim']]
```

[ ]:

	sent1	sent2	frech_sim	sim
9	a man is playing a guitar	a man is playing a trumpet	0.985250	1.714
10	a man is playing a guitar	a man is playing a trumpet	0.985250	1.714
12	a man is cycling	a man is talking	0.983160	0.600
15	a man is playing a guitar	a man is playing a keyboard	0.954951	1.800
28	a man is speaking	a man is cooking	0.903790	0.800
126	someone is drawing	someone is dancing	0.901419	0.300
149	a man is praying	a man is dancing	0.985135	0.750
157	a man is spitting	a man is talking	0.717959	0.800
159	a man is dancing	a man is thinking	0.857412	1.200
612	a group of kids are having a jumping contest	a group of kids are having a sleepover	0.983003	1.200
675	there is no maximum	there is no quarantine period	0.964640	0.000
769	this is a great one	this is a difficult one	0.826343	1.000
841	you should do it	you should never do it	0.771359	1.000
1182	kl shares higher at midafternoon	kl shares lower at midmorning	0.642191	1.600
1208	china stocks close higher on friday	china stocks close higher on wednesday	0.642191	1.800

[ ]:
```
# False Negatives: where the frechet similarity thinks they are not similar but they are
sts_test[(sts_test['frech_sim'] >= 5) & (sts_test['sim'] >= 4)][['sent1', 'sent2', 'frech_sim', 'sim']]
```

[ ]:

	sent1	\
768	well i wouldnt put it on my cv	
947	second comes hp 27 percent with 29 billion up …	
1069	tomorrow at the mission inn i have the opportu…	
1299	georgian pms ally to become president exit poll	
1322	croatia begins countdown to historic eu entry	
1335	uks expremier margaret thatcher dies at 87	

	sent2	frech_sim	sim
768	i wouldnt put this job on my resume	5.657111	4.0
947	hp fell to second place with server sales grow…	6.330658	4.4
1069	nan	5.017543	4.0
1299	ally of georgias billionaire pm to be presiden…	5.704584	4.6
1322	croatia countdowns to joining eu	7.541843	4.8
1335	former british pm margaret thatcher dies	5.056745	4.0