```
 1 "C:\Users\Eric Liao\Anaconda3\envs\python2\python.exe" "C
   :\Program Files (x86)\JetBrains\PyCharm Community Edition
   2016.3.2\helpers\pydev\pydevconsole.py" 53816 53817
 2 Python 2.7.12 |Anaconda 2.4.1 (64-bit)| (default, Jun 29
   2016, 11:07:13) [MSC v.1500 64 bit (AMD64)]
 3 Type "copyright", "credits" or "license" for more
   information.
 4
 5 IPython 4.0.1 -- An enhanced Interactive Python.
 6 ?         -> Introduction and overview of IPython's
   features.
 7 %quickref -> Quick reference.
 8 help      -> Python's own help system.
 9 object?   -> Details about 'object', use 'object??' for
   extra details.
10 PyDev console: using IPython 4.0.1
11
12 import sys; print('Python %s on %s' % (sys.version, sys.
   platform))
13 sys.path.extend(['C:\\Users\\Eric Liao\\Desktop\\Intro
   machine learning\\ml_projects', 'C:/Users/Eric Liao/
   Desktop/Intro machine learning/ml_projects'])
14
15 Python 2.7.12 |Anaconda 2.4.1 (64-bit)| (default, Jun 29
   2016, 11:07:13) [MSC v.1500 64 bit (AMD64)] on win32
16 In[2]: from sklearn.feature_extraction.text import
   CountVectorizer
17 In[3]: vectorizer = CountVectorizer(min_df=1)
18 In[4]: print(vetorizer)
19 Traceback (most recent call last):
20   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\site
   -packages\IPython\core\interactiveshell.py", line 3066, in
    run_code
21     exec(code_obj, self.user_global_ns, self.user_ns)
22   File "<ipython-input-4-76895eac4259>", line 1, in <
   module>
23     print(vetorizer)
24 NameError: name 'vetorizer' is not defined
25 In[5]: print(vectorizer)
26 CountVectorizer(analyzer=u'word', binary=False,
   decode_error=u'strict',
27         dtype=<type 'numpy.int64'>, encoding=u'utf-8',
   input=u'content',
28         lowercase=True, max_df=1.0, max_features=None,
   min_df=1,
29         ngram_range=(1, 1), preprocessor=None, stop_words=
   None,
```

```
30          strip_accents=None, token_pattern=u'(?u)\\b\\w\\w
   +\\b',
31          tokenizer=None, vocabulary=None)
32 In[6]: content = ["How to format my hard disk", " Hard
   disk format problems "]
33 In[7]: X = vectorizer.fit_transform(content)
34 In[8]: vectorizer.get_feature_names()
35 Out[8]: [u'disk', u'format', u'hard', u'how', u'my', u'
   problems', u'to']
36 In[9]: print(X.toarray().transpose())
37 [[1 1]
38  [1 1]
39  [1 1]
40  [1 0]
41  [1 0]
42  [0 1]
43  [1 0]]
44 In[10]: posts = [open(os.path.join(DIR, f)).read() for f
   in os.listdir(DIR)]
45 Traceback (most recent call last):
46   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\site
   -packages\IPython\core\interactiveshell.py", line 3066, in
    run_code
47     exec(code_obj, self.user_global_ns, self.user_ns)
48   File "<ipython-input-10-7b68b5c4249c>", line 1, in <
   module>
49     posts = [open(os.path.join(DIR, f)).read() for f in os
   .listdir(DIR)]
50 NameError: name 'os' is not defined
51 In[11]: import os
52 In[12]: posts = [open(os.path.join(DIR, f)).read() for f
   in os.listdir(DIR)]
53 Traceback (most recent call last):
54   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\site
   -packages\IPython\core\interactiveshell.py", line 3066, in
    run_code
55     exec(code_obj, self.user_global_ns, self.user_ns)
56   File "<ipython-input-12-7b68b5c4249c>", line 1, in <
   module>
57     posts = [open(os.path.join(DIR, f)).read() for f in os
   .listdir(DIR)]
58 NameError: name 'DIR' is not defined
59 In[13]: DIR = "C:\Users\Eric Liao\Desktop\Intro machine
   learning\ml_projects"
60 In[14]: DIR = "C:\Users\Eric Liao\Desktop\Intro machine
   learning\ml_projects\DIR"
61 In[15]: posts = [open(os.path.join(DIR, f)).read() for f
```

```
in os.listdir(DIR)]
In[16]: vectorizer = CountVectorizer(min_df=1)
In[17]: X_train = vectorizer.fit_transform(posts)
In[18]: num_samples, num_features = X_train.shape
In[19]: print("#samples: %d, #features: %d" % (
num_samples, num_features))
#samples: 5, #features: 24
In[20]: print(vectorizer.get_feature_names())
[u'about', u'actually', u'can', u'contains', u'databases
', u'get', u'huge', u'images', u'imaging', u'interesting
', u'is', u'it', u'learning', u'machine', u'most', u'much
', u'not', u'permanently', u'post', u'safe', u'store', u'
stuff', u'this', u'toy']
In[21]: new_post = "imaging databases"
In[22]: new_post_vec = vectorizer.transform([new_post])
In[23]: print(new_post_vec)
  (0, 4)    1
  (0, 8)    1
In[24]: print(new_post_vec.toarray())
[[0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]
In[25]: import scipy as sp
In[26]: def dist_raw(v1, v2):
   ...:     delta = v1 - v2
   ...:     return sp.linalg.norm(delta.toarray())
   ...:
In[27]: import sys
In[28]: best_doc = None
In[29]: best_dist = sys.maxint
In[30]: best_i = NOne
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
site-packages\IPython\core\interactiveshell.py", line
3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-30-e71d9e0d6e10>", line 1, in <
module>
    best_i = NOne
NameError: name 'NOne' is not defined
In[31]: best_i = None
In[32]: for i in range(0, num_samples):
   ...:     post = posts[i]
   ...:     if post == new_post:
   ...:         continue
   ...:     post_vec = X_train.getrow(i)
   ...:     d = dist(post_vec, new_post_vec)
   ...:
Traceback (most recent call last):
```

```
100   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
101     exec(code_obj, self.user_global_ns, self.user_ns)
102   File "<ipython-input-32-2a30a99d4ed1>", line 6, in <
    module>
103     d = dist(post_vec, new_post_vec)
104 NameError: name 'dist' is not defined
105 In[33]: for i in range(0, num_samples):
106    ...:     post = posts[i]
107    ...:     if post == new_post:
108    ...:         continue
109    ...:     post_vec = X_train.getrow(i)
110    ...:     d = dist(post_vec, new_post_vec)
111    ...:     print "== post %i with dist=%.2f: %s" % (i, d
    , post)
112    ...:     if d < best_dist:
113    ...:         best_dist = d
114    ...:         best_i = i
115    ...:
116 Traceback (most recent call last):
117   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
118     exec(code_obj, self.user_global_ns, self.user_ns)
119   File "<ipython-input-33-28d8ef2dbc93>", line 6, in <
    module>
120     d = dist(post_vec, new_post_vec)
121 NameError: name 'dist' is not defined
122 In[34]: for i in range(0, num_samples):
123    ...:     post = posts[i]
124    ...:     if post == new_post:
125    ...:         continue
126    ...:     post_vec = X_train.getrow(i)
127    ...:     d = dist_raw(post_vec, new_post_vec)
128    ...:     print "== post %i with dist=%.2f: %s" % (i, d
    , post)
129    ...:     if d < best_dist:
130    ...:         best_dist = d
131    ...:         best_i = i
132    ...:
133 == post 0 with dist=4.00: This is a toy post about
    machine learning. Actually, it contains not much
    interesting stuff.
134 == post 1 with dist=1.73: Imaging databases can get huge.
135 == post 2 with dist=2.00: Most imaging databases safe
    images permanently.
```

```
136 == post 3 with dist=1.41: Imaging databases store images.
137 == post 4 with dist=5.10: Imaging databases store images
    . Imaging databases store images. Imaging databases store
     images.
138 In[35]: print(X_train.getrow(3).toarray())
139 [[0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0]]
140 In[36]: print(X_train.getrow(4).toarray())
141 [[0 0 0 0 3 0 0 3 3 0 0 0 0 0 0 0 0 0 0 3 0 0 0]]
142 In[37]: # Normalizing the word count vectors
143 In[38]: def dist_norm(v1, v2):
144    ...:     v1_normalized = v1 / sp.linalg.norm(v1.
    toarray())
145    ...:     v1_normalized = v2 / sp.linalg.norm(v2.
    toarray())
146    ...:     delta = v1_normalized - v2_normalized
147    ...:     return sp.linalg.norm(delta.toarray())
148    ...:
149 In[39]: for i in range(0, num_samples):
150    ...:     post = posts[i]
151    ...:     if post == new_post:
152    ...:         continue
153    ...:     post_vec = X_train.getrow(i)
154    ...:     d = dist_raw(post_vec, new_post_vec)
155    ...:     print "== post %i with dist=%.2f: %s" % (i, d
    , post)
156    ...:     if d < best_dist:
157    ...:         best_dist = d
158    ...:         best_i = i
159    ...:
160 == post 0 with dist=4.00: This is a toy post about
    machine learning. Actually, it contains not much
    interesting stuff.
161 == post 1 with dist=1.73: Imaging databases can get huge.
162 == post 2 with dist=2.00: Most imaging databases safe
    images permanently.
163 == post 3 with dist=1.41: Imaging databases store images.
164 == post 4 with dist=5.10: Imaging databases store images
    . Imaging databases store images. Imaging databases store
     images.
165 In[40]: # Removing less important words
166 In[41]: vectorizer = CountVectorizer(min_df=1, stop_words
    ='english')
167 In[42]: sorted(vectorizer.get_stop_words())[0:20]
168 Out[42]:
169 ['a',
170  'about',
171  'above',
```

```
172  'across',
173  'after',
174  'afterwards',
175  'again',
176  'against',
177  'all',
178  'almost',
179  'alone',
180  'along',
181  'already',
182  'also',
183  'although',
184  'always',
185  'am',
186  'among',
187  'amongst',
188  'amoungst']
189 In[43]: import nltk
190 In[44]: import nltk.stem
191 Backend Qt4Agg is interactive backend. Turning
    interactive mode on.
192 In[45]: s = nltk.stem.SnowballStemmer('english')
193 In[46]: s.stem("graphics")
194 Out[46]: u'graphic'
195 In[47]: s.stem("imaging")
196 Out[47]: u'imag'
197 In[48]: s.stem("image")
198 Out[48]: u'imag'
199 In[49]: english_stemmer = nltk.stem.SnowballStemmer('
    english')
200 In[50]: class StemmedCountVectorizer(CountVectorizer):
201    ...:     def build_analyzer(self):
202    ...:         analyzer = super(StemmedCountVectorizer,
    self).build_analyzer()
203    ...:         return lambda doc: (english_stemmer.stem(
    w) for w in analyzer(doc))
204    ...:
205 In[51]: vectorizer = StemmedCountVectorizer(min_df=1,
    stop_words='english')
206 In[52]: X_train = vectorizer.fit_transform(posts)
207 In[53]: print(vectorizer.get_feature_names())
208 [u'actual', u'contain', u'databas', u'huge', u'imag', u'
    interest', u'learn', u'machin', u'perman', u'post', u'
    safe', u'store', u'stuff', u'toy']
209 In[54]: num_samples, num_features = X_train.shape
210    ...: print("#samples: %d, #features: %d" % (
    num_samples, num_features))
```

```
211    ...:
212 #samples: 5, #features: 14
213 In[55]: for i in range(0, num_samples):
214    ...:     post = posts[i]
215    ...:     if post == new_post:
216    ...:         continue
217    ...:     post_vec = X_train.getrow(i)
218    ...:     d = dist_norm(post_vec, new_post_vec)
219    ...:     print "== post %i with dist=%.2f: %s" % (i, d
    , post)
220    ...:     if d < best_dist:
221    ...:         best_dist = d
222    ...:         best_i = i
223    ...:
224 Traceback (most recent call last):
225   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
226     exec(code_obj, self.user_global_ns, self.user_ns)
227   File "<ipython-input-55-c5d5d2cd3140>", line 6, in <
    module>
228     d = dist_norm(post_vec, new_post_vec)
229   File "<ipython-input-38-8656066cdc9b>", line 4, in
    dist_norm
230     delta = v1_normalized - v2_normalized
231 NameError: global name 'v2_normalized' is not defined
232 In[56]: def dist_norm(v1, v2):
233    ...:     v1_normalized = v1 / sp.linalg.norm(v1.
    toarray())
234    ...:     v1_normalized = v2 / sp.linalg.norm(v2.
    toarray())
235    ...:     delta = v1_normalized - v2_normalized
236    ...:     return sp.linalg.norm(delta.toarray())
237    ...:
238 In[57]: def dist_norm(v1, v2):
239    ...:     v1_normalized = v1 / sp.linalg.norm(v1.
    toarray())
240    ...:     v2_normalized = v2 / sp.linalg.norm(v2.
    toarray())
241    ...:     delta = v1_normalized - v2_normalized
242    ...:     return sp.linalg.norm(delta.toarray())
243    ...:
244 In[58]: for i in range(0, num_samples):
245    ...:     post = posts[i]
246    ...:     if post == new_post:
247    ...:         continue
248    ...:     post_vec = X_train.getrow(i)
```

```
249    ...:     d = dist_norm(post_vec, new_post_vec)
250    ...:     print "== post %i with dist=%.2f: %s" % (i, d
    , post)
251    ...:     if d < best_dist:
252    ...:         best_dist = d
253    ...:         best_i = i
254    ...:
255 Traceback (most recent call last):
256   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
257     exec(code_obj, self.user_global_ns, self.user_ns)
258   File "<ipython-input-58-c5d5d2cd3140>", line 6, in <
    module>
259     d = dist_norm(post_vec, new_post_vec)
260   File "<ipython-input-57-6aaf61685bcf>", line 4, in
    dist_norm
261     delta = v1_normalized - v2_normalized
262   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\scipy\sparse\compressed.py", line 371, in
    __sub__
263     raise ValueError("inconsistent shapes")
264 ValueError: inconsistent shapes
265 In[59]: new_post_vec = vectorizer.transform([new_post])
266 In[60]: print(new_post_vec)
267   (0, 2)    1
268   (0, 4)    1
269 In[61]: for i in range(0, num_samples):
270    ...:     post = posts[i]
271    ...:     if post == new_post:
272    ...:         continue
273    ...:     post_vec = X_train.getrow(i)
274    ...:     d = dist_norm(post_vec, new_post_vec)
275    ...:     print "== post %i with dist=%.2f: %s" % (i, d
    , post)
276    ...:     if d < best_dist:
277    ...:         best_dist = d
278    ...:         best_i = i
279    ...:
280 == post 0 with dist=1.41: This is a toy post about
    machine learning. Actually, it contains not much
    interesting stuff.
281 == post 1 with dist=0.61: Imaging databases can get huge.
282 == post 2 with dist=0.63: Most imaging databases safe
    images permanently.
283 == post 3 with dist=0.52: Imaging databases store images.
284 == post 4 with dist=0.52: Imaging databases store images
```

```
. Imaging databases store images. Imaging databases store images.
In[62]:
In[62]: # term frequency - inverse document frequency (TF-IDF)
In[63]: import scipy as sp
In[64]: def tfidf(term, doc, docset):
   ...:     tf = float(doc.count(term)) / sum(doc.count(w) for w in docset)
   ...:     idf = math.log(float(len(docset))) / (len([doc for doc in docset if term in doc]))
   ...:     return tf * idf
   ...:
In[65]: a, abb, abc = ["a"], ["a", "b", "b"], ["a", "b", "c"]
In[66]: a
Out[66]: ['a']
In[67]: D = [a, abb, abc]
In[68]: D
Out[68]: [['a'], ['a', 'b', 'b'], ['a', 'b', 'c']]
In[69]: abb
Out[69]: ['a', 'b', 'b']
In[70]: print(tfidf("a", a, D))
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\site-packages\IPython\core\interactiveshell.py", line 3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-70-d9939f3ea873>", line 1, in <module>
    print(tfidf("a", a, D))
  File "<ipython-input-64-cb87d3f6a6a9>", line 2, in tfidf
    tf = float(doc.count(term)) / sum(doc.count(w) for w in docset)
ZeroDivisionError: float division by zero
In[71]: print(tfidf("b", abb, D))
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\site-packages\IPython\core\interactiveshell.py", line 3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-71-2c78b9a59f88>", line 1, in <module>
    print(tfidf("b", abb, D))
  File "<ipython-input-64-cb87d3f6a6a9>", line 2, in tfidf
```

```
317     tf = float(doc.count(term)) / sum(doc.count(w) for w
    in docset)
318 ZeroDivisionError: float division by zero
319 In[72]: sum(a.count(w) for w in D)
320 Out[72]: 0
321 In[73]: D = ['a', 'abb', 'abc']
322 In[74]: sum(a.count(w) for w in D)
323 Out[74]: 1
324 In[75]: print(tfidf("a", a, D))
325    ...:
326 Traceback (most recent call last):
327   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
328     exec(code_obj, self.user_global_ns, self.user_ns)
329   File "<ipython-input-75-d9939f3ea873>", line 1, in <
    module>
330     print(tfidf("a", a, D))
331   File "<ipython-input-64-cb87d3f6a6a9>", line 3, in
    tfidf
332     idf = math.log(float(len(docset))) / (len([doc for
    doc in docset if term in doc]))
333 NameError: global name 'math' is not defined
334 In[76]: import math
335 In[77]: print(tfidf("a", a, D))
336 0.366204096223
337 In[78]: D = [a, abb, abc]
338    ...:
339 In[79]: for w in D:
340    ...:     print w
341    ...:
342 ['a']
343 ['a', 'b', 'b']
344 ['a', 'b', 'c']
345 In[80]: a.count('a')
346 Out[80]: 1
347 In[81]: a.count(['a'])
348 Out[81]: 0
349 In[82]:
350 In[82]: # Getting test data to evaluate our ideas on
351 In[83]:
352 In[83]: import sklearn.datasets
353 In[84]: MLCOMP_DIR = r"C:\Users\Eric Liao\Desktop\Intro
    machine learning\ml_projects\clustering_post\data\dataset
    -379-20news-18828_NMJWK"
354 In[85]: data = sklearn.datasets.load_mlcomp("20news-18828
    ", mlcomp_root = MLCOMP_DIR)
```

```
In[86]: print(data.filenames)
[ 'C:\\Users\\Eric Liao\\Desktop\\Intro machine learning
    \\ml_projects\\clustering_post\\data\\dataset-379-20news-
    18828_NMJWK\\379\\raw\\comp.graphics\\1190-38614'
 'C:\\Users\\Eric Liao\\Desktop\\Intro machine learning\\
    ml_projects\\clustering_post\\data\\dataset-379-20news-
    18828_NMJWK\\379\\raw\\comp.graphics\\1383-38616'
 'C:\\Users\\Eric Liao\\Desktop\\Intro machine learning\\
    ml_projects\\clustering_post\\data\\dataset-379-20news-
    18828_NMJWK\\379\\raw\\alt.atheism\\487-53344'
 ...,
 'C:\\Users\\Eric Liao\\Desktop\\Intro machine learning\\
    ml_projects\\clustering_post\\data\\dataset-379-20news-
    18828_NMJWK\\379\\raw\\rec.sport.hockey\\10215-54303'
 'C:\\Users\\Eric Liao\\Desktop\\Intro machine learning\\
    ml_projects\\clustering_post\\data\\dataset-379-20news-
    18828_NMJWK\\379\\raw\\sci.crypt\\10799-15660'
 'C:\\Users\\Eric Liao\\Desktop\\Intro machine learning\\
    ml_projects\\clustering_post\\data\\dataset-379-20news-
    18828_NMJWK\\379\\raw\\comp.os.ms-windows.misc\\2732-
    10871']
In[87]: print(len(data.filenames))
18828
In[88]: data.target_names
Out[88]:
['alt.atheism',
 'comp.graphics',
 'comp.os.ms-windows.misc',
 'comp.sys.ibm.pc.hardware',
 'comp.sys.mac.hardware',
 'comp.windows.x',
 'misc.forsale',
 'rec.autos',
 'rec.motorcycles',
 'rec.sport.baseball',
 'rec.sport.hockey',
 'sci.crypt',
 'sci.electronics',
 'sci.med',
 'sci.space',
 'soc.religion.christian',
 'talk.politics.guns',
 'talk.politics.mideast',
 'talk.politics.misc',
 'talk.religion.misc']
In[89]: train_data = sklearn.datasets.load_mlcomp("20news
    -18828", "train", ml_root=MLCOMP_DIR)
```

```
388 Traceback (most recent call last):
389   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
390     exec(code_obj, self.user_global_ns, self.user_ns)
391   File "<ipython-input-89-441adf8affbc>", line 1, in <
    module>
392     train_data = sklearn.datasets.load_mlcomp("20news-
    18828", "train", ml_root=MLCOMP_DIR)
393   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\datasets\mlcomp.py", line 62, in
    load_mlcomp
394     raise ValueError("MLCOMP_DATASETS_HOME env variable
    is undefined")
395 ValueError: MLCOMP_DATASETS_HOME env variable is
    undefined
396 In[90]: train_data = sklearn.datasets.load_mlcomp("20news
    -18828", "train", mlcomp_root=MLCOMP_DIR)
397 In[91]: print(len(train_data.filenames))
398 13180
399 In[92]: test_data = sklearn.datasets.load_mlcomp("20news-
    18828",
400    ...: "test", mlcomp_root=MLCOMP_DIR)
401 In[93]: print(len(test_data.filenames))
402 5648
403 In[94]: groups = ['comp.graphics', 'comp.os.ms-windows.
    misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.ma c.
    hardware', 'comp.windows.x', 'sci.space']
404 In[95]: train_data = sklearn.datasets.load_mlcomp("20news
    -18828", "train", mlcomp_root=MLCOMP_DIR, categories=
    groups)
405 In[96]: print(len(train_data.filenames))
406 3414
407 In[97]:
408 In[97]: # Clustering posts
409 In[98]:
410 In[98]: vectorizer = StemmedTfidVectorizer(min_df=10,
    max_df=0.5, stop_words='english', charset_error='ignore')
411 Traceback (most recent call last):
412   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
413     exec(code_obj, self.user_global_ns, self.user_ns)
414   File "<ipython-input-98-583ff4b27c97>", line 1, in <
    module>
415     vectorizer = StemmedTfidVectorizer(min_df=10, max_df=
    0.5, stop_words='english', charset_error='ignore')
```

```
NameError: name 'StemmedTfidVectorizer' is not defined
In[99]: vectorizer = StemmedTfidfVectorizer(min_df=10,
    max_df=0.5, stop_words='english', charset_error='ignore')
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-99-3ccfebc3c0e2>", line 1, in <
    module>
    vectorizer = StemmedTfidfVectorizer(min_df=10, max_df
    =0.5, stop_words='english', charset_error='ignore')
NameError: name 'StemmedTfidfVectorizer' is not defined
In[100]: vectorizer = StemmedTfidfVectorizer(min_df=10,
    max_df=0.5, stop_words='english', charset_error='ignore')
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-100-3ccfebc3c0e2>", line 1, in <
    module>
    vectorizer = StemmedTfidfVectorizer(min_df=10, max_df
    =0.5, stop_words='english', charset_error='ignore')
NameError: name 'StemmedTfidfVectorizer' is not defined
In[101]: class StemmedTfidfVectorizer(TfidfVectorizer):
    ...:     def build_analyzer(self):
    ...:         analyzer = super(TfidfVectorizer, self).
    build_analyzer()
    ...:         return lambda doc: (english_stemmer.stem
    (w) for w in analyzer(doc))
    ...:
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-101-e39ab45fdaa4>", line 1, in <
    module>
    class StemmedTfidfVectorizer(TfidfVectorizer):
NameError: name 'TfidfVectorizer' is not defined
In[102]: from sklearn.feature_extraction.text import
    TfidfVectorizer
In[103]: class StemmedTfidfVectorizer(TfidfVectorizer):
    ...:     def build_analyzer(self):
    ...:         analyzer = super(TfidfVectorizer, self).
    build_analyzer()
```

```
    ...:         return lambda doc: (english_stemmer.stem
(w) for w in analyzer(doc))
    ...:
In[104]: vectorizer = StemmedTfidfVectorizer(min_df=10,
max_df=0.5, stop_words='english', charset_error='ignore')
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
site-packages\IPython\core\interactiveshell.py", line
3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-104-3ccfebc3c0e2>", line 1, in <
module>
    vectorizer = StemmedTfidfVectorizer(min_df=10, max_df
=0.5, stop_words='english', charset_error='ignore')
TypeError: __init__() got an unexpected keyword argument
 'charset_error'
In[105]: vectorizer = StemmedTfidfVectorizer(min_df=10,
max_df=0.5, stop_words='english', decode_error='ignore')
In[106]: vectorized = vectorizer.fit_transform(dataset.
data)
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
site-packages\IPython\core\interactiveshell.py", line
3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-106-ad4894d9170d>", line 1, in <
module>
    vectorized = vectorizer.fit_transform(dataset.data)
NameError: name 'dataset' is not defined
In[107]: vectorized = vectorizer.fit_transform(train_data
)
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
site-packages\IPython\core\interactiveshell.py", line
3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-107-fd4cf555680c>", line 1, in <
module>
    vectorized = vectorizer.fit_transform(train_data)
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
site-packages\sklearn\feature_extraction\text.py", line
1305, in fit_transform
    X = super(TfidfVectorizer, self).fit_transform(
raw_documents)
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
site-packages\sklearn\feature_extraction\text.py", line
834, in fit_transform
```

```
472     "max_df corresponds to < documents than min_df")
473 ValueError: max_df corresponds to < documents than min_df
474 In[108]: vectorizer = StemmedTfidfVectorizer(min_df=0.5,
    max_df=10, stop_words='english', decode_error='ignore')
475 In[109]: vectorized = vectorizer.fit_transform(train_data
    )
476 Traceback (most recent call last):
477   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
478     exec(code_obj, self.user_global_ns, self.user_ns)
479   File "<ipython-input-109-fd4cf555680c>", line 1, in <
    module>
480     vectorized = vectorizer.fit_transform(train_data)
481   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    1305, in fit_transform
482     X = super(TfidfVectorizer, self).fit_transform(
    raw_documents)
483   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    838, in fit_transform
484     max_features)
485   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    733, in _limit_features
486     raise ValueError("After pruning, no terms remain. Try
     a lower"
487 ValueError: After pruning, no terms remain. Try a lower
    min_df or a higher max_df.
488 In[110]: vectorizer = StemmedTfidfVectorizer(min_df=0.5,
    max_df=1.0, stop_words='english', decode_error='ignore')
489 In[111]: vectorized = vectorizer.fit_transform(train_data
    )
490 Traceback (most recent call last):
491   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
492     exec(code_obj, self.user_global_ns, self.user_ns)
493   File "<ipython-input-111-fd4cf555680c>", line 1, in <
    module>
494     vectorized = vectorizer.fit_transform(train_data)
495   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    1305, in fit_transform
496     X = super(TfidfVectorizer, self).fit_transform(
    raw_documents)
```

```
497   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    838, in fit_transform
498     max_features)
499   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    733, in _limit_features
500     raise ValueError("After pruning, no terms remain. Try
     a lower"
501 ValueError: After pruning, no terms remain. Try a lower
    min_df or a higher max_df.
502 In[112]: vectorizer = StemmedTfidfVectorizer(min_df=0.6,
    max_df=1.0, stop_words='english', decode_error='ignore')
503 In[113]: vectorized = vectorizer.fit_transform(train_data
    )
504 Traceback (most recent call last):
505   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
506     exec(code_obj, self.user_global_ns, self.user_ns)
507   File "<ipython-input-113-fd4cf555680c>", line 1, in <
    module>
508     vectorized = vectorizer.fit_transform(train_data)
509   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    1305, in fit_transform
510     X = super(TfidfVectorizer, self).fit_transform(
    raw_documents)
511   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    838, in fit_transform
512     max_features)
513   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    733, in _limit_features
514     raise ValueError("After pruning, no terms remain. Try
     a lower"
515 ValueError: After pruning, no terms remain. Try a lower
    min_df or a higher max_df.
516 In[114]: vectorizer = StemmedTfidfVectorizer(min_df=0.6,
    max_df=0.95, stop_words='english', decode_error='ignore')
517 In[115]: vectorized = vectorizer.fit_transform(train_data
    )
518 Traceback (most recent call last):
519   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
    3066, in run_code
```

```
520     exec(code_obj, self.user_global_ns, self.user_ns)
521   File "<ipython-input-115-fd4cf555680c>", line 1, in <
    module>
522     vectorized = vectorizer.fit_transform(train_data)
523   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    1305, in fit_transform
524     X = super(TfidfVectorizer, self).fit_transform(
    raw_documents)
525   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    838, in fit_transform
526     max_features)
527   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\sklearn\feature_extraction\text.py", line
    733, in _limit_features
528     raise ValueError("After pruning, no terms remain. Try
     a lower"
529 ValueError: After pruning, no terms remain. Try a lower
    min_df or a higher max_df.
530 In[116]: vectorizer = StemmedTfidfVectorizer(max_df=0.95
    , stop_words='english', decode_error='ignore')
531 In[117]: vectorized = vectorizer.fit_transform(train_data
    )
532 In[118]: num_samples, num_features = vectorized.shape
533 In[119]: print("#samples: %d, #features: %d" % (
    num_samples, num_features))
534 #samples: 5, #features: 5
535 In[120]: vectorizer = StemmedTfidfVectorizer(min_df=0.1,
    max_df=0.5, stop_words='english', decode_error='ignore')
536 In[121]: vectorized = vectorizer.fit_transform(train_data
    )
537 In[122]: print("#samples: %d, #features: %d" % (
    num_samples, num_features))
538 #samples: 5, #features: 5
539 In[123]: print(vectorized)
540   (0, 4)    1.0
541   (1, 0)    1.0
542   (2, 3)    1.0
543   (3, 1)    1.0
544   (4, 2)    1.0
545 In[124]: print(len(train_data.filenames))
546 3414
547 In[125]: dataset.data
548 Traceback (most recent call last):
549   File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
    site-packages\IPython\core\interactiveshell.py", line
```

```
3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-125-96b0c1b6f0e5>", line 1, in <
module>
    dataset.data
NameError: name 'dataset' is not defined
In[126]: datasets.data
Traceback (most recent call last):
  File "C:\Users\Eric Liao\Anaconda3\envs\python2\lib\
site-packages\IPython\core\interactiveshell.py", line
3066, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)
  File "<ipython-input-126-4c8ede9d9fe8>", line 1, in <
module>
    datasets.data
NameError: name 'datasets' is not defined

```