Autoimmune Tweets using the mostly preprocessed file from R and testing on Lemmatized Tweets with 8 categories of autoimmune diseases¶

Those being: 0:Leukemia, 1: Fibromyalgia, 2:Kidney Disease, 3: Celiac Disease, 4: MS, 5: Hashimoto, 6: RA, 7: Chron's Disease

Tweets were taken from respective diseases in early December 2019 from 13 to 119 tweets for each disease, as many as were found that weren't mostly marketing, using '' treatment' in the search

In [1]:

```
%matplotlib inline
import pandas as pd
import matplotlib.pyplot as plt
from textblob import TextBlob
import sklearn
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report, f1_score, accuracy_score, confusion
_matrix


np.random.seed(47)
```

In [2]:

```
reviews = pd.read_csv('LemmaPythonRead.csv', encoding = 'unicode_escape')
#the encoding needed for python3 handling nonASCII chars
```

In [3]:

```
reviews.head()
```

Out[3]:

	LemmatizedTweets	StemmedTweets	AutoImmuneDisorder
0	unknown research unknownresearch the center fo...	unknown research\r\nunknownresearch\r\nthe cen...	Celiac_Disease
1	lynn barter abc mc lbarter · dec reply to thre...	lynn barter abc mc\r\nlbarter\r\n·\r\ndec \r\n...	Celiac_Disease
2	theona layne theonawrites · dec unknown diseas...	theona layne\r\ntheonawrites\r\n·\r\ndec \r\nu...	Celiac_Disease

	LemmatizedTweets	StemmedTweets	AutoImmuneDisorder
3	bob simonoff simonoffbob · dec there be eviden...	bob simonoff\r\nsimonoffbob\r\n·\r\ndec \r\nth...	Celiac_Disease
4	gfdenver gfdenver · nov hm interest research n...	gfdenver\r\ngfdenver\r\n·\r\nnov \r\nhm intere...	Celiac_Disease

In [4]:

```
reviews.tail()
```

Out[4]:

	LemmatizedTweets	StemmedTweets	AutoImmuneDisorder
502	pharmabot thepharmabot · nov codessly effectiv...	pharmabot\r\nthepharmabot\r\n·\r\nnov \r\ncode...	Leukemia_Disease
503	wcm lymphoma wcmclymphoma · dec select initial...	wcm lymphoma\r\nwcmclymphoma\r\n·\r\ndec \r\ns...	Leukemia_Disease
504	medivizor medivizor · dec cope with cml check ...	medivizor\r\nmedivizor\r\n·\r\ndec \r\ncoping ...	Leukemia_Disease
505	abi brokenleadheart · dec reply to rickyspurs ...	abi\r\nbrokenleadheart\r\n·\r\ndec \r\nreplyin...	Leukemia_Disease
506	brooke xbrooke · dec reply to itsjojosiwa dear...	brooke\r\n\r\n\r\nxbrooke\r\n·\r\ndec \r\nrepl...	Leukemia_Disease

In [5]:

```
reviews.shape
```

Out[5]:

```
(507, 3)
```

In [6]:

```
reviews = reviews.reindex(np.random.permutation(reviews.index))


print(reviews.head())

print(reviews.tail())
```

```
                                      LemmatizedTweets  \
407  medivizor medivizor · nov cope with cml check ...
196  medical news bulletin mednewsbulletin · jun a ...
359  drtharanga kumari wickramasooriya drtharanga ·...
39   nola unknown unknowndiary · sep reply to nolan...
245  christine blome blomechristine · jan our new t...


                                         StemmedTweets AutoImmuneDisorder
407  medivizor\r\nmedivizor\r\n·\r\nnov \r\ncoping ...   Leukemia_Disease
196  medical news bulletin\r\nmednewsbulletin\r\n·\...       Fibromyalgia
359  drtharanga kumari wickramasooriya\r\ndrtharang...     Kidney_Disease
39   nola unknown\r\nunknowndiary\r\n·\r\nsep  \r\n...     Celiac_Disease
245  christine blome\r\nblomechristine\r\n·\r\njan ...         MS_Disease
                                      LemmatizedTweets  \
72   r unknownunknown runknownunknown · h chronic o...
264  lorilynn lorilynn · nov multiple unknown be a ...
327  unknown guild□ theunknownguild · nov fridayfin...
390  drug topic drugtopics · dec the fda have appro...
```

```
135  fms news bot fmsbot · nov unknown treatment ma...

                                           StemmedTweets AutoImmuneDisorder
72   r unknownunknown\r\nrunknownunknown\r\n·\r\nh\...  Hashimoto_Disease
264  lorilynn\r\nlorilynn\r\n·\r\nnov \r\nmultiple ...         MS_Disease
327  unknown guild\r\ntheunknownguild\r\n·\r\nnov ...         MS_Disease
390  drug topics\r\ndrugtopics\r\n·\r\ndec \r\nthe ...   Leukemia_Disease
135  fms news bot\r\nfmsbot\r\n·\r\nnov \r\nunknown...       Fibromyalgia
```

In [7]:

```
reviews.groupby('AutoImmuneDisorder').describe()
```

Out[7]:

	LemmatizedTweets				**StemmedTweets**			
	count	**unique**	**top**	**freq**	**count**	**unique**	**top**	**freq**
AutoImmuneDisorder								
Celiac_Disease	50	50	blake parson blakepparsons · dec need help for...	1	50	50	np agarwal\r\nnpgrwl\r\n·\r\nnov \r\ntreatment...	1
Chron_Disease	19	19	thomas and ethel bakery thomasandethel · nov r...	1	19	19	tyler daniel\r\ntylerdaniel\r\n·\r\naug \r\nb...	1
Fibromyalgia	99	95	fibro bloggers fibrobloggers · nov unknown tre...	2	99	95	chronic disease coalition\r\nchronicrights\r\n...	2
Hashimoto_Disease	30	29	colorado natural med drgravesco · dec naturopa...	2	30	29	colorado natural med\r\ndrgravesco\r\n·\r\ndec...	2
Kidney_Disease	43	43	stock shark stocksharks · dec today announce p...	1	43	43	eclinic diagnostics\r\neclinicnigeria\r\nyour...	1
Leukemia_Disease	119	116	medivizor medivizor · nov cope with cml check ...	3	119	116	medivizor\r\nmedivizor\r\n·\r\nnov \r\ncoping ...	3
MS_Disease	119	119	uonresearch uonresearch · jan fund signal new ...	1	119	119	multiple sclerosis\r\nunknownbio\r\n·\r\nnov \...	1
RA_Disease	28	28	gse health blog gsehealth · sep what be the tr...	1	28	28	frontiers medicine\r\nfrontmedicine\r\n·\r\noc...	1

In [8]:

```
reviews.groupby('AutoImmuneDisorder').describe()
```

Out[8]:

	LemmatizedTweets				StemmedTweets			
	count	unique	top	freq	count	unique	top	freq
AutoImmuneDisorder								
Celiac_Disease	50	50	blake parson blakepparsons · dec need help for...	1	50	50	np agarwal\r\nnpgrwl\r\n·\r\nnov \r\ntreatment...	1
Chron_Disease	19	19	thomas and ethel bakery thomasandethel · nov r...	1	19	19	tyler daniel\r\ntylerdaniel\r\n·\r\naug \r\nb...	1
Fibromyalgia	99	95	fibro bloggers fibrobloggers · nov unknown tre...	2	99	95	chronic disease coalition\r\nchronicrights\r\n...	2
Hashimoto_Disease	30	29	colorado natural med drgravesco · dec naturopa...	2	30	29	colorado natural med\r\ndrgravesco\r\n·\r\ndec...	2
Kidney_Disease	43	43	stock shark stocksharks · dec today announce p...	1	43	43	eclinic diagnostics\r\neclinicnigeria\r\nyour...	1
Leukemia_Disease	119	116	medivizor medivizor · nov cope with cml check ...	3	119	116	medivizor\r\nmedivizor\r\n·\r\nnov \r\ncoping ...	3
MS_Disease	119	119	uonresearch uonresearch · jan fund signal new ...	1	119	119	multiple sclerosis\r\nunknownbio\r\n·\r\nnov \...	1
RA_Disease	28	28	gse health blog gsehealth · sep what be the tr...	1	28	28	frontiers medicine\r\nfrontmedicine\r\n·\r\noc...	1

In [9]:

```
reviews['length'] = reviews['LemmatizedTweets'].map(lambda text: len(text))
print(reviews.head())
```

```
                                          LemmatizedTweets  \
407  medivizor medivizor · nov cope with cml check ...
196  medical news bulletin mednewsbulletin · jun a ...
359  drtharanga kumari wickramasooriya drtharanga ·...
39   nola unknown unknowndiary · sep reply to nolan...
245  christine blome blomechristine · jan our new t...

                                             StemmedTweets AutoImmuneDisorder  \
407  medivizor\r\nmedivizor\r\n·\r\nnov \r\ncoping ...   Leukemia_Disease
196  medical news bulletin\r\nmednewsbulletin\r\n·\...       Fibromyalgia
359  drtharanga kumari wickramasooriya\r\ndrtharang...     Kidney_Disease
39   nola unknown\r\nunknowndiary\r\n·\r\nsep  \r\n...     Celiac_Disease
245  christine blome\r\nblomechristine\r\n·\r\njan ...         MS_Disease

     length
407     126
196     245
```

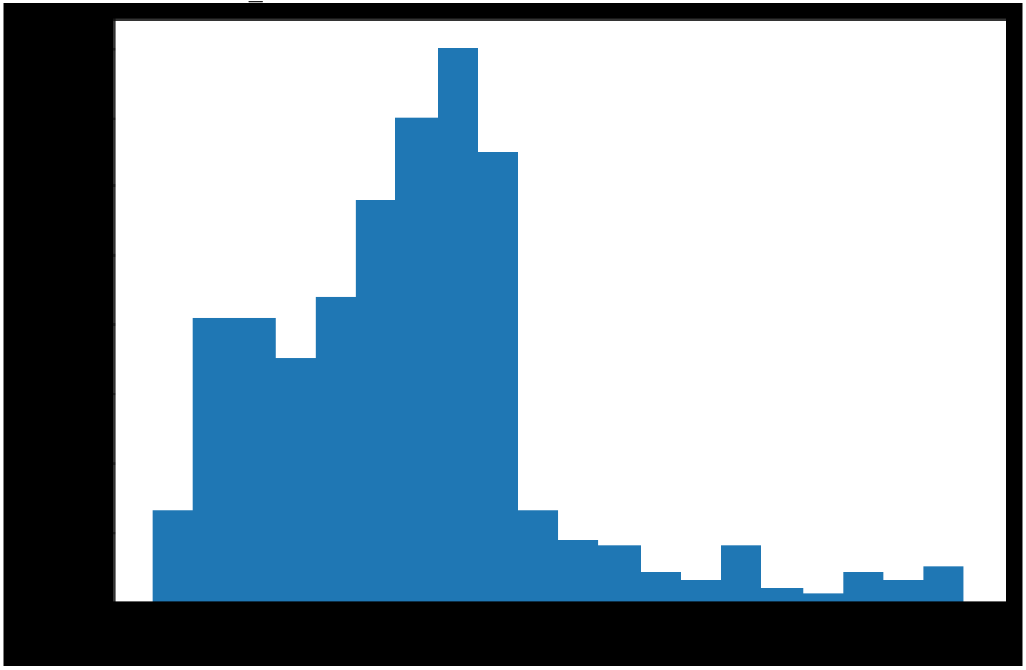
```
359      312
39       319
245      196
```

In [10]:

```
reviews.length.plot(bins=20, kind='hist')
```

Out[10]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x18f59ff4128>
```



In [11]:

```
reviews.length.describe()
```

Out[11]:

```
count    507.000000
mean     243.998028
std       92.843285
min       77.000000
25%      175.500000
50%      244.000000
75%      287.000000
max      604.000000
Name: length, dtype: float64
```

In [12]:

```
print(list(reviews.LemmatizedTweets[reviews.length > 500].index)) #near the max for le
ngth of LemmatizedTweets

print(list(reviews.AutoImmuneDisorder[reviews.length > 500]))
```

```
[75, 432, 105, 104, 58, 145, 26, 82, 109, 111, 99, 167, 149]
['Hashimoto_Disease', 'Leukemia_Disease', 'Fibromyalgia', 'Fibromyalgia', 'Hashimoto_D
isease', 'Fibromyalgia', 'Celiac_Disease', 'Chron_Disease', 'Fibromyalgia', 'Fibromyal
gia', 'Fibromyalgia', 'Fibromyalgia', 'Fibromyalgia']
```

In [13]:

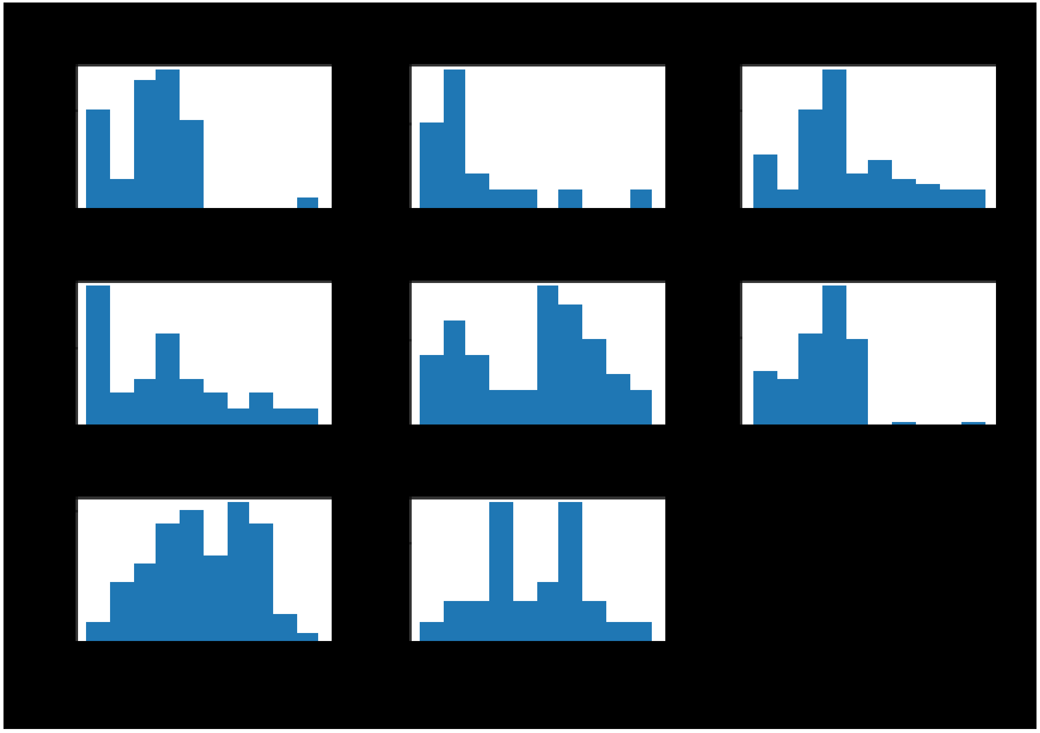
```
%%time

reviews.hist(column='length', by='AutoImmuneDisorder', bins=10)
```

```
Wall time: 516 ms
```

Out[13]:

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000018F5A1780B8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000018F5A1C4E48>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000018F5A1FA438>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000018F5A22C9E8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000018F5A25DF28>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000018F5A29A518>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000018F5A2CCAC8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000018F5A3090F0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000018F5A309128>]],
      dtype=object)
```



In [14]:

```
def split_into_tokens(review):


    #review = unicode(review, 'iso-8859-1')# in python 3 the default of str() previously python2 as unicode() is utf-8

    return TextBlob(review).words
```

In [15]:

```
reviews.LemmatizedTweets.head().apply(split_into_tokens)
```

Out[15]:

```
407    [medivizor, medivizor, ·, nov, cope, with, cml...
196    [medical, news, bulletin, mednewsbulletin, ·, ...
359    [drtharanga, kumari, wickramasooriya, drtharan...
39     [nola, unknown, unknowndiary, ·, sep, reply, t...
245    [christine, blome, blomechristine, ·, jan, our...
Name: LemmatizedTweets, dtype: object
```

In [16]:

```
TextBlob("hello world, how is it going?").tags
```

Out[16]:

```
[('hello', 'JJ'),
 ('world', 'NN'),
 ('how', 'WRB'),
 ('is', 'VBZ'),
 ('it', 'PRP'),
 ('going', 'VBG')]
```

In [17]:

```
import nltk

nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\m\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[17]:

```
True
```

In [18]:

```
from nltk.corpus import stopwords


stop = stopwords.words('english')

stop = stop + [u'a',u'b',u'c',u'd',u'e',u'f',u'g',u'h',u'i',u'j',u'k',u'l',u'm',u'n',u
'o',u'p',u'q',u'r',u's',u't',u'v',u'w',u'x',u'y',u'z']
```

In [19]:

```
def split_into_lemmas(review):

    #review = unicode(review, 'iso-8859-1')

    review = review.lower()

    #review = unicode(review, 'utf8').lower()

    #review = str(review).lower()

    words = TextBlob(review).words

    # for each word, take its "base form" = lemma

    return [word.lemma for word in words if word not in stop]
```

```
reviews.LemmatizedTweets.head().apply(split_into_lemmas)
```

Out[19]:

```
407    [medivizor, medivizor, ·, nov, cope, cml, chec...
196    [medical, news, bulletin, mednewsbulletin, ·, ...
359    [drtharanga, kumari, wickramasooriya, drtharan...
39     [nola, unknown, unknowndiary, ·, sep, reply, n...
245    [christine, blome, blomechristine, ·, jan, new...
Name: LemmatizedTweets, dtype: object
```

In [20]:

```
%%time

bow_transformer = CountVectorizer(analyzer=split_into_lemmas).fit(reviews['LemmatizedT
weets'])

print(len(bow_transformer.vocabulary_))
```

```
4208
Wall time: 734 ms
```

In [21]:

```
review4 = reviews['LemmatizedTweets'][42]

print(review4)
```

```
purna purnamusic · jun gluten shouldn □ t be so painful no sleep night two advice try
antihistamine ginger tea ibuprofen and activate charcoal over the last hour unknown tr
eatment
```

In [22]:

```
bow4 = bow_transformer.transform([review4])

print(bow4)
```

```
  (0, 49)	1
  (0, 81)	1
  (0, 214)	1
  (0, 628)	1
  (0, 1458)	1
  (0, 1476)	1
  (0, 1665)	1
  (0, 1871)	1
  (0, 2113)	1
  (0, 2191)	1
  (0, 2713)	1
  (0, 2848)	1
  (0, 3129)	1
  (0, 3130)	1
  (0, 3473)	1
  (0, 3702)	1
  (0, 3845)	1
  (0, 3873)	1
  (0, 3890)	1
  (0, 3944)	1
  (0, 4199)	1
  (0, 4206)	1
```

In [23]:

```
%%time

reviews_bow = bow_transformer.transform(reviews['LemmatizedTweets'])

print('sparse matrix shape:', reviews_bow.shape)

print('number of non-zeros:', reviews_bow.nnz)

print('sparsity: %.2f%%' % (100.0 * reviews_bow.nnz / (reviews_bow.shape[0] * reviews_bow.shape[1])))
```

```
sparse matrix shape: (507, 4208)
number of non-zeros: 11902
sparsity: 0.56%
Wall time: 781 ms
```

In [24]:

```
# Split/splice into training ~ 80% and testing ~ 20%

reviews_bow_train = reviews_bow[:400]

reviews_bow_test = reviews_bow[400:]

reviews_sentiment_train = reviews['AutoImmuneDisorder'][:400]

reviews_sentiment_test = reviews['AutoImmuneDisorder'][400:]


print(reviews_bow_train.shape)

print(reviews_bow_test.shape)
```

```
(400, 4208)
(107, 4208)
```

In [25]:

```
%time review_sentiment = MultinomialNB().fit(reviews_bow_train, reviews_sentiment_train)
```

```
Wall time: 78.1 ms
```

In [26]:

```
print('predicted:', review_sentiment.predict(bow4)[0])

print('expected:', reviews.AutoImmuneDisorder[42])
```

```
predicted: Celiac_Disease
expected: Celiac_Disease
```

In [27]:

```
predictions = review_sentiment.predict(reviews_bow_test)

print(predictions)
```

```
['Fibromyalgia' 'Fibromyalgia' 'MS_Disease' 'Leukemia_Disease'
 'MS_Disease' 'Fibromyalgia' 'Leukemia_Disease' 'Kidney_Disease'
 'Hashimoto_Disease' 'Fibromyalgia' 'Fibromyalgia' 'Leukemia_Disease'
 'Fibromyalgia' 'Fibromyalgia' 'MS_Disease' 'MS_Disease' 'MS_Disease'
 'Fibromyalgia' 'Fibromyalgia' 'Leukemia_Disease' 'Leukemia_Disease'
 'Fibromyalgia' 'Fibromyalgia' 'Leukemia_Disease' 'Fibromyalgia'
 'Celiac_Disease' 'Leukemia_Disease' 'Fibromyalgia' 'Leukemia_Disease'
 'Leukemia_Disease' 'Fibromyalgia' 'Leukemia_Disease' 'Leukemia_Disease'
 'MS_Disease' 'MS_Disease' 'Fibromyalgia' 'Leukemia_Disease' 'MS_Disease'
 'MS_Disease' 'Fibromyalgia' 'Hashimoto_Disease' 'MS_Disease' 'MS_Disease'
 'MS_Disease' 'MS_Disease' 'Leukemia_Disease' 'MS_Disease' 'MS_Disease'
 'Celiac_Disease' 'Fibromyalgia' 'Fibromyalgia' 'Fibromyalgia'
 'MS_Disease' 'Leukemia_Disease' 'Fibromyalgia' 'MS_Disease'
 'Leukemia_Disease' 'MS_Disease' 'Leukemia_Disease' 'Kidney_Disease'
 'MS_Disease' 'Fibromyalgia' 'Fibromyalgia' 'MS_Disease'
 'Leukemia_Disease' 'Leukemia_Disease' 'Fibromyalgia' 'Fibromyalgia'
 'Leukemia_Disease' 'Fibromyalgia' 'Celiac_Disease' 'MS_Disease'
 'Fibromyalgia' 'MS_Disease' 'Hashimoto_Disease' 'Leukemia_Disease'
 'MS_Disease' 'MS_Disease' 'Celiac_Disease' 'MS_Disease' 'Fibromyalgia'
 'MS_Disease' 'MS_Disease' 'Fibromyalgia' 'Leukemia_Disease'
 'Leukemia_Disease' 'MS_Disease' 'RA_Disease' 'Hashimoto_Disease'
 'Celiac_Disease' 'MS_Disease' 'Hashimoto_Disease' 'Celiac_Disease'
 'Fibromyalgia' 'Fibromyalgia' 'Celiac_Disease' 'MS_Disease'
 'Fibromyalgia' 'Hashimoto_Disease' 'Celiac_Disease' 'MS_Disease'
 'Fibromyalgia' 'Fibromyalgia' 'MS_Disease' 'MS_Disease'
 'Leukemia_Disease' 'Fibromyalgia']
```

In [28]:

```
print('accuracy', accuracy_score(reviews_sentiment_test, predictions))
print('confusion matrix\n', confusion_matrix(reviews_sentiment_test, predictions))
print('(row=expected, col=predicted)')
```

```
accuracy 0.6635514018691588
confusion matrix
 [[ 2  0  1  0  1  0  3  0]
 [ 2  0  0  0  0  0  2  1]
 [ 0  0 23  0  0  0  2  0]
 [ 2  0  3  6  0  0  0  0]
 [ 1  0  0  0  1  1  1  0]
 [ 0  0  0  0  0 21  1  0]
 [ 1  0  2  0  0  1 18  0]
 [ 0  0  5  0  0  0  6  0]]
(row=expected, col=predicted)
```

In [29]:

```
print(classification_report(reviews_sentiment_test, predictions))
#The F1 score can be interpreted as a weighted average of the precision and recall,
#where an F1 score reaches its best value at 1 and worst score at 0.
```

```
                   precision    recall  f1-score   support

   Celiac_Disease       0.25      0.29      0.27         7
    Chron_Disease       0.00      0.00      0.00         5
     Fibromyalgia       0.68      0.92      0.78        25
Hashimoto_Disease       1.00      0.55      0.71        11
   Kidney_Disease       0.50      0.25      0.33         4
 Leukemia_Disease       0.91      0.95      0.93        22
       MS_Disease       0.55      0.82      0.65        22
       RA_Disease       0.00      0.00      0.00        11

         accuracy                           0.66       107
        macro avg       0.49      0.47      0.46       107
     weighted avg       0.60      0.66      0.61       107

c:\users\m\anaconda2\envs\python36\lib\site-packages\sklearn\metrics\classification.py
:1437: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
```

In [40]:

```
def predict_review(new_review):

    new_sample = bow_transformer.transform([new_review])

    p = np.around(review_sentiment.predict_proba(new_sample), decimals=2)

    print(new_review, '\t', p, '\tMax: ', np.max(p), '\n')
```

The respective probabilities correspond to those diseases alphebatized as
[[1-Celiac Disease, 2-Chron's Disease, 3-Fibromyalgia, 4-Hashimoto, 5-Kidney Disease, 6-Leukemia, 7-Multiple Sclerosis, 8-Rheumatoid Arthritis]]

In [39]:

```
predict_review('sick. pain. sleepless. anxious.')


predict_review('digestive. hungry.')


predict_review('bruising. sleepy. tired. headache.')

predict_review('energy. crazy. manic. depressed. angry.')
```

```
sick. pain. sleepless. anxious.    [[0.01 0.   0.88 0.01 0.01 0.01 0.07 0.01]]       M
ax:  0.88

digestive. hungry.         [[0.11 0.03 0.19 0.05 0.1  0.24 0.24 0.04]]       Max:  0.24

bruising. sleepy. tired. headache.         [[0.09 0.08 0.39 0.05 0.09 0.13 0.13 0.05]]
        Max:  0.39

energy. crazy. manic. depressed. angry.    [[0.1  0.05 0.17 0.06 0.1  0.22 0.24 0.06]]
        Max:  0.24
```

In [ ]: