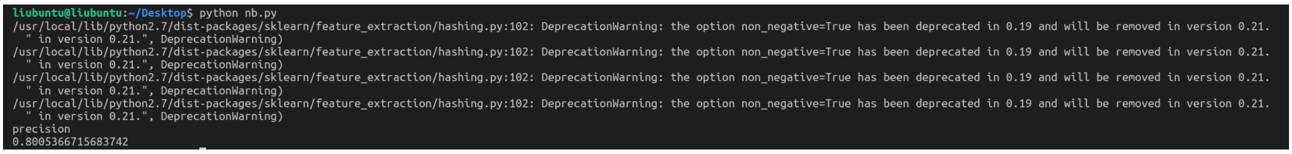
Homework2：Naive Bayes

代码及原理：

```
from sklearn.datasets import fetch_20newsgroups
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import HashingVectorizer
from sklearn.datasets import fetch_20newsgroups_vectorized
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics

def main():
    categories = ['comp.graphics',
    'comp.os.ms-windows.misc',
    'comp.sys.ibm.pc.hardware',
    'comp.sys.mac.hardware',
    'comp.windows.x',
    'rec.autos',
    'rec.motorcycles',
    'rec.sport.baseball',
    'rec.sport.hockey',
    'sci.crypt',
    'sci.electronics',
    'sci.med',
    'sci.space',
    'misc.forsale',
    'talk.politics.misc',
    'talk.politics.guns',
    'talk.politics.mideast',
    'talk.religion.misc',
    'alt.atheism',
    'soc.religion.christian'];

   #下载数据
    newsgroup_train = fetch_20newsgroups(subset = 'train',categories = categories);
    newsgroups_test = fetch_20newsgroups(subset = 'test',
                                  categories = categories);

  #将数据向量化
    vectorizer = HashingVectorizer(stop_words = 'english',non_negative = True,
                             n_features = 10000)
    fea_train = vectorizer.fit_transform(newsgroup_train.data)
    fea_test = vectorizer.fit_transform(newsgroups_test.data);
    #创建朴素贝叶斯分类器
    clf = MultinomialNB(alpha = 0.01)
    clf.fit(fea_train,newsgroup_train.target);
    #用朴素贝叶斯分类器预测测试集
    pred = clf.predict(fea_test);
   #计算结果
    calculate_result(newsgroups_test.target,pred);
```

```
def calculate_result(actual,pred):
    m_precision = metrics.precision_score(actual,pred,average="weighted");
    print "precision"
    print m_precision

if __name__ == "__main__":
    main()
```

根据运行结果观察到准确率为：0.8005366715683742



```
liubuntu@liubuntu:~/Desktop$ python nb.py
/usr/local/lib/python2.7/dist-packages/sklearn/feature_extraction/hashing.py:102: DeprecationWarning: the option non_negative=True has been deprecated in 0.19 and will be removed in version 0.21.
  " in version 0.21.", DeprecationWarning)
/usr/local/lib/python2.7/dist-packages/sklearn/feature_extraction/hashing.py:102: DeprecationWarning: the option non_negative=True has been deprecated in 0.19 and will be removed in version 0.21.
  " in version 0.21.", DeprecationWarning)
/usr/local/lib/python2.7/dist-packages/sklearn/feature_extraction/hashing.py:102: DeprecationWarning: the option non_negative=True has been deprecated in 0.19 and will be removed in version 0.21.
  " in version 0.21.", DeprecationWarning)
/usr/local/lib/python2.7/dist-packages/sklearn/feature_extraction/hashing.py:102: DeprecationWarning: the option non_negative=True has been deprecated in 0.19 and will be removed in version 0.21.
  " in version 0.21.", DeprecationWarning)
precision
0.8005366715683742
```