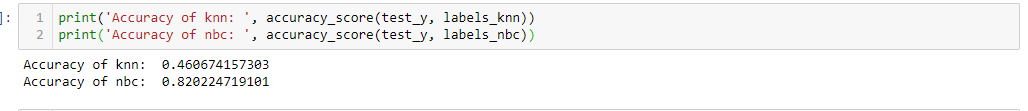
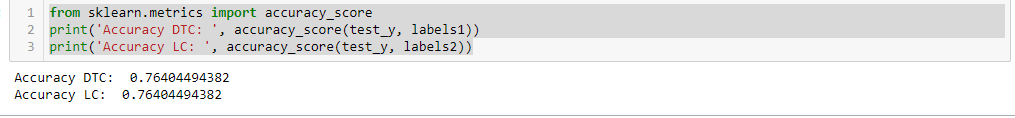
Assignment 2: Natural language of processing





As it is shown in the pictures, The Top three accuracy of my result are from NBC, DTC, and LC

**Code:**

**KNN and NBC:**

from sklearn.neighbors import KNeighborsClassifier

from sklearn.naive\_bayes import MultinomialNB

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.feature\_extraction.text import CountVectorizer

corpus = open('badges.data').read()

data = corpus.split('\n')

data.remove(data[-1])

data.remove(data[0])

X\_label,Y\_value=[],[]

for row in data:

label = row[:1]

value = row[2:]

X\_label.append(label)

Y\_value.append(value)

vec = CountVectorizer(max\_features = 20)

matrix\_Y = vec.fit\_transform(Y\_value)

from sklearn.feature\_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer()

matrix\_Y = vec.fit\_transform(Y\_value)

train\_x,test\_x,train\_y,test\_y = train\_test\_split(matrix\_Y,X\_label,shuffle=True,train\_size = 0.7)

knnc = KNeighborsClassifier(n\_neighbors = 5)

nbc = MultinomialNB()

knn.fit(train\_x, train\_y)

nbc.fit(train\_x, train\_y)

labels\_knn = knn.predict(test\_x)

labels\_nbc = nbc.predict(test\_x)

print('Accuracy of knn: ', accuracy\_score(test\_y, labels\_knn))

print('Accuracy of nbc: ', accuracy\_score(test\_y, labels\_nbc))

**DTC and LC**

**from sklearn.tree import DecisionTreeClassifier**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.linear\_model import SGDClassifier**

**from sklearn.feature\_extraction.text import TfidfVectorizer**

**from sklearn.feature\_extraction.text import CountVectorizer**

**corpus = open('badges.data').read()**

**X, y = [], []**

**for row in data:**

**label = row[:1]**

**name = row[2:]**

**X.append(name)**

**y.append(label)**

**from sklearn.feature\_extraction.text import CountVectorizer**

**vec = CountVectorizer(max\_features = 20)**

**matrix\_X = vec.fit\_transform(X)**

**train\_x, test\_x, train\_y, test\_y = train\_test\_split(matrix\_X, y, shuffle = True, train\_size = 0.7)**

**dtc = DecisionTreeClassifier()**

**lc = SGDClassifier()**

**lc.fit(train\_x, train\_y)**

**dtc.fit(train\_x, train\_y)**

**labels1 = dtc.predict(test\_x)**

**labels2 = lc.predict(test\_x)**

**from sklearn.metrics import accuracy\_score**

**print('Accuracy DTC: ', accuracy\_score(test\_y, labels1))**

**print('Accuracy LC: ', accuracy\_score(test\_y, labels2))**