Assignment 2 - DS4Biz Y63

TextScraping_Classification

Team Detail

Team Name: Platinum Bread

Student 1

Student ID: 61070315
Student Full Name: ภูธเนศ สุวรรณเคาม

Student 2

Student ID: 61070326
Student Full Name: สุวิภา นันชัย

แหล่งข้อมูล

http://www.it.kmitl.ac.th/~teerapong/news_archive/index.html

In [1]:
```
import pandas as pd
import numpy as np
import re
import bs4
import requests
import string
import operator
import urllib.request
import nltk
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction import text
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression, SGDClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import classification_report
from bs4 import BeautifulSoup
from pprint import pprint
```

Part 1: Data Collection

ทำการดึงข้อมูล

In [2]:
```
response = requests.get('http://www.it.kmitl.ac.th/~teerapong/news_archive/index.html')
html_page = bs4.BeautifulSoup(response.content, 'lxml')
print(html_page)
```

```
<!DOCTYPE html>
<html lang="en">
<head>
<title>Online News Archive</title>
<meta charset="utf-8"/>
<meta content="width=device-width, initial-scale=1" name="viewport"/>
<meta content="noindex" name="robots"/>
<meta content="news,articles,news" name="keywords"/>
<meta content="Breaking News | International Headlines" property="og:title"/>
<meta content="News Archive" property="og:site_name"/>
<meta content="Latest news and more from the definitive brand of quality news." property="og:description"/>
<link href="css/bootstrap.min.css" rel="stylesheet"/>
<script src="js/jquery-3.2.1.slim.min.js"></script>
<script src="js/popper.min.js"></script>
<script src="js/tether.min.js"></script>
<script src="js/jquery-3.2.1.slim.min.js"></script>
<script src="js/popper.min.js"></script>
<script src="js/tether.min.js"></script>
<script src="js/bootstrap.min.js"></script>
<style>
                        .main{ padding: 0; text-align: center;}
                        .footer{ padding: 6px;text-align: center; margin-top: 1em; }

                        h1
                        {
                                font-size: 180%;
                                margin-top: 15px;
                                margin-bottom: 15px;
                        }
                        ul {list-style-type: none;}
                        li { margin-top: 5px; }
        </style>
</head>
<body>
<div class="container" style="margin-top: 2em;">
<div class="main">
<img alt="banner" src="images/banner.jpg" width="500"/>
<h1>News Article Archive</h1>
<p>Archive of all news headlines and stories, organised per month.</p>
<ul>
```

```
<li>Articles – <a href="month-jan-2017.html">January</a> [118]</li>
<li>Articles – <a href="month-feb-2017.html">February</a> [124]</li>
<li>Articles – <a href="month-mar-2017.html">March</a> [116]</li>
<li>Articles – <a href="month-apr-2017.html">April</a> [118]</li>
<li>Articles – <a href="month-may-2017.html">May</a> [115]</li>
<li>Articles – <a href="month-jun-2017.html">June</a> [115]</li>
<li>Articles – <a href="month-jul-2017.html">July</a> [122]</li>
<li>Articles – <a href="month-aug-2017.html">August</a> [116]</li>
<li>Articles – <a href="month-sep-2017.html">September</a> [113]</li>
<li>Articles – <a href="month-oct-2017.html">October</a> [124]</li>
<li>Articles – <a href="month-nov-2017.html">November</a> [122]</li>
<li>Articles – <a href="month-dec-2017.html">December</a> [115]</li>
</ul>
</div>
<div class="footer">
<span><a href="#">Terms &amp;amp; Conditions</a> | <a href="#">Privacy Policy</a> | <a href="#">Cookie Information</a> </span><br/>
<span>© <span class="thisyear">2019-2020</span> – Original rights holders</span>
</div>
</div>
</body>
</html>
```

ดึงข้อมูลลิ้งค์ที่จะเข้าถึงเนื้อหา

In [3]:
```
tags = html_page.select('li>a')
extract_href = [] #list เก็บข้อมูล href เพื่อที่จะเข้าถึงหน้า Article
for i in tags :
    extract_href.append(i['href'])
```

In [4]:
```
tags
```

Out[4]:
```
[<a href="month-jan-2017.html">January</a>,
 <a href="month-feb-2017.html">February</a>,
 <a href="month-mar-2017.html">March</a>,
 <a href="month-apr-2017.html">April</a>,
 <a href="month-may-2017.html">May</a>,
 <a href="month-jun-2017.html">June</a>,
 <a href="month-jul-2017.html">July</a>,
 <a href="month-aug-2017.html">August</a>,
 <a href="month-sep-2017.html">September</a>,
 <a href="month-oct-2017.html">October</a>,
 <a href="month-nov-2017.html">November</a>,
 <a href="month-dec-2017.html">December</a>]
```

In [5]:
```
extract_href
```

Out[5]:
```
['month-jan-2017.html',
 'month-feb-2017.html',
 'month-mar-2017.html',
 'month-apr-2017.html',
 'month-may-2017.html',
 'month-jun-2017.html',
 'month-jul-2017.html',
 'month-aug-2017.html',
 'month-sep-2017.html',
 'month-oct-2017.html',
 'month-nov-2017.html',
 'month-dec-2017.html']
```

In [6]:
```
link_article = [] #list เก็บข้อมูลของ article ไว้ใช้ดึงเนื้อหาข่าว
extract_category = [] #list เก็บข้อมูล category
for i in extract_href :
    response = requests.get('http://www.it.kmitl.ac.th/~teerapong/news_archive/' + i )
    html_page = bs4.BeautifulSoup(response.content, 'lxml')
    month = html_page.select('tr>td>a')
    category = html_page.select('tr>td.category')

    for i in month: #Loop ดึงข้อมูล
        link_article.append(i['href'])

    for i in category: #Loop ดึงข้อมูลที่ category
        if i.text.strip() == 'N/A': #ลบค่าที่เป็น N/A
            del i
        else:
            extract_category.append(i.text.strip()) #.stip() เพื่อกำจัดอักขระ "\xa0"ออก
```

In [7]:
```
link_article
```

Out[7]:
```
['article-jan-0418.html',
 'article-jan-0027.html',
 'article-jan-0631.html',
 'article-jan-2105.html',
 'article-jan-3300.html',
 'article-jan-4187.html',
 'article-jan-1974.html',
 'article-jan-3666.html',
 'article-jan-2629.html',
 'article-jan-2415.html',
 'article-jan-4210.html',
 'article-jan-4789.html',
 'article-jan-3452.html',
 'article-jan-2428.html',
 'article-jan-4766.html',
 'article-jan-2595.html',
 'article-jan-2935.html',
 'article-jan-0578.html',
 'article-jan-3023.html',
 'article-jan-2356.html'
```

In [8]:
```
extract_category
```

Out[8]:
```
['technology',
 'business',
 'technology',
 'business',
 'sport',
 'sport',
 'sport',
 'sport',
 'technology',
 'technology',
 'sport',
 'business',
 'business',
 'technology',
 'technology',
```

```
 'technology',
 'business',
 'business',
 'technology',
 'sport'
```

In [9]:
```
all_article_heading = []
for i in link_article:
    article = ''
    response = requests.get('http://www.it.kmitl.ac.th/~teerapong/news_archive/' + i)
    html_page = bs4.BeautifulSoup(response.content, 'lxml')
    selector_h2 = html_page.select('h2')
    article += selector_h2[0].text
    selector_p = html_page.select('p')
    for i in selector_p[:-1]:
        article += i.text+' '
    all_article_heading.append(article)
```

In [10]:
```
all_article_heading
```

Out[10]:
```
['21st-Century Sports: How Digital Technology Is Changing the Face Of The Sporting Industry The sporting industry has come a long way since the ‘60s. It has carved out for itself a niche with its roots so deep that I cannot fathom the sports industry showing any sign of decline any time soon - or later.  The reason can be found in this seemingly subtle difference - other industries have customers; the sporting industry has fans. Vivek Ranadivé, leader of the ownership group of the NBA’s Sacramento Kings, explained it beautifully, “Fans will paint their face purple, fans will evangelize. ... Every other CEO in every business is dying to be in our position – they’re dying to have fans.“  While fan passion alone could almost certainly keep the industry going, leagues and sporting franchises have decided not to rest on their laurels. The last few years have seen the steady introduction of technology into the world of sports - amplifying fans’ appreciation of games, enhancing athletes’ public profiles and informing their training methods, even influencing how contests are waged.  Also, digital technology in particular has helped to create an alternative source of revenue, besides the games themselves - corporate sponsorship. They achieved this by capitalizing on the ardor of their customer base - sorry, fan base. ',
 'Asian quake hits European sharesAsian quake hits European shares  Shares in Europe\'s leading reinsurers and travel firms have fallen as the scale of the damage wrought by tsunamis across south Asia has become apparent.  More than 23,000 people have been killed following a massive underwater earthquake and many of the worst hit areas are popular tourist destinations. Reinsurance firms such as Swiss Re and Munich Re lost value as investors worried about rebuilding costs. But the disaster has little impact on stock markets in the US and Asia.  Currencies including the Thai baht and Indonesian rupiah weakened as analysts warned that economic growth may slow. "It came at the worst possible time," said Hans Goetti, a Singapore-based fund manager. "The impact on the tourist industry is pretty devastating, especially in Thailand." Travel-related shares dropped in Europe, with companies such as Germany\'s TUI and Lufthansa and France\'s Club Mediterranne sliding. Insurers and reinsurance firms were also under pressure in Europe.  Shares in Munich Re and Swiss Re - the world\'s two biggest reinsurers - both fell 1
```

In [11]:
```
all_article_content = []
for i in link_article:
    article = ''
    response = requests.get('http://www.it.kmitl.ac.th/~teerapong/news_archive/' + i)
    html_page = bs4.BeautifulSoup(response.content, 'lxml')
    selector_p = html_page.select('p')
    for i in selector_p[:-1]:
        article += i.text+' '
    all_article_content.append(article)
```

In [12]:
```
all_article_content
```

Out[12]:
```
[' The sporting industry has come a long way since the ‘60s. It has carved out for itself a niche with its roots so deep that I cannot fathom the sports industry showing any sign of decline any time soon - or later.  The reason can be found in this seemingly subtle difference - other industries have customers; the sporting industry has fans. Vivek Ranadivé, leader of the ownership group of the NBA’s Sacramento Kings, explained it beautifully, “Fans will paint their face purple, fans will evangelize. ... Every other CEO in every business is dying to be in our position – they’re dying to have fans.“  While fan passion alone could almost certainly keep the industry going, leagues and sporting franchises have decided not to rest on their laurels. The last few years have seen the steady introduction of technology into the world of sports - amplifying fans’ appreciation of games, enhancing athletes’ public profiles and informing their training methods, even influencing how contests are waged.  Also, digital technology in particular has helped to create an alternative source of revenue, besides the games themselves - corporate sponsorship. They achieved this by capitalizing on the ardor of their customer base - sorry, fan base. ',
 'Asian quake hits European shares  Shares in Europe\'s leading reinsurers and travel firms have fallen as the scale of the damage wrought by tsunamis across south Asia has become apparent.  More than 23,000 people have been killed following a massive underwater earthquake and many of the worst hit areas are popular tourist destinations. Reisurance firms such as Swiss Re and Munich Re lost value as investors worried about rebuilding costs. But the disaster has little impact on stock markets in the US and Asia.  Currencies including the Thai baht and Indonesian rupiah weakened as analysts warned that economic growth may slow. "It came at the worst possible time," said Hans Goetti, a Singapore-based fund manager. "The impact on the tourist industry is pretty devastating, especially in Thailand." Travel-related shares dropped in Europe, with companies such as Germany\'s TUI and Lufthansa and France\'s Club Mediterranne sliding. Insurers and reinsurance firms were also under pressure in Europe.  Shares in Munich Re and Swiss Re - the world\'s two biggest reinsurers - both fell 1.7% as the market speculated about the cost of rebuilding in Asia. Zurich Financial, Allianz and Axa also suffered a decline in value.  However, their losses we
```

Save to Text Flie

save category.txt

In [13]:
```
with open('target/category.txt', 'a', encoding = 'utf-8') as f:
    for i in extract_category:
        f.write(i +'\n')
```

save AllArticles_HeadingPlusContent.txt

In [14]:
```
with open('datastore/AllArticles_HeadingPlusContent.txt', 'a', encoding = 'utf-8') as f:
    for i in all_article_heading:
        f.write(i +'\n')
```

save AllArticles_OnlyContent.txt

In [15]:
```
with open('datastore/AllArticles_OnlyContent.txt', 'a', encoding = 'utf-8') as f:
    for i in all_article_content:
        f.write(i +'\n')
```

Part 2: Text Classification

Read Text File : AllArticles_HeadingPlusContent.txt

In [16]:
```
with open('target/category.txt', 'r', encoding = 'utf-8') as f:
    category_data = f.read().splitlines()
category_data
```

Out[16]:
```
['technology',
 'business',
 'technology',
 'business',
 'sport',
 'sport',
 'sport',
 'sport',
 'technology',
 'technology',
 'sport',
```

```
 'business',
 'business',
 'technology',
 'technology',
 'technology',
 'business',
 'business',
 'technology',
 'sport'
```

In [17]:
```
#เรียกไฟล์ AllArticles_HeadingPlusContent.txt
with open('datastore/AllArticles_OnlyContent.txt', 'r', encoding = 'utf-8') as f:
    content_data = f.read().splitlines()
content_data
```

Out[17]:
```
[' The sporting industry has come a long way since the ‘60s. It has carved out for itself a niche with its roots so deep that I cannot fathom the sports industry showing any sign of decline any time soon - or later.  The reason can be found in this seemingly subtle difference - other industries have customers; the sporting industry has fans. Vivek Ranadivé, leader of the ownership group of the NBA’s Sacramento Kings, explained it beautifully, “Fans will paint their face purple, fans will evangelize. ... Every other CEO in every business is dying to be in our position – they’re dying to have fans.“  While fan passion alone could almost certainly keep the industry going, leagues and sporting franchises have decided not to rest on their laurels. The last few years have seen the steady introduction of technology into the world of sports - amplifying fans’ appreciation of games, enhancing athletes’ public profiles and informing their training methods, even influencing how contests are waged.  Also, digital technology in particular has helped to create an alternative source of revenue, besides the games themselves - corporate sponsorship. They achieved this by capitalizing on the ardor of their customer base - sorry, fan base. ',
 'Asian quake hits European shares  Shares in Europe\'s leading reinsurers and travel firms have fallen as the scale of the damage wrought by tsunamis across south Asia has become apparent.  More than 23,000 people have been killed following a massive underwater earthquake and many of the worst hit areas are popular tourist destinations. Reisurance firms such as Swiss Re and Munich Re lost value as investors worried about rebuilding costs. But the disaster has little impact on stock markets in the US and Asia.  Currencies including the Thai baht and Indonesian rupiah weakened as analysts warned that economic growth may slow. "It came at the worst possible time," said Hans Goetti, a Singapore-based fund manager. "The impact on the tourist industry is pretty devastating, especially in Thailand." Travel-related shares dropped in Europe, with companies such as Germany\'s TUI and Lufthansa and France\'s Club Mediterranne sliding. Insurers and reinsurance firms were also under pressure in Europe.  Shares in Munich Re and Swiss Re - the world\'s two biggest reinsurers - both fell 1.7% as the market speculated about the cost of rebuilding in Asia.  Zurich Financial, Allianz and Axa also suffered a decline in value.  However, their losses we
```

In [18]:
```
df = pd.DataFrame(list(zip(content_data, category_data)), columns=['content','category'])
df
```

Out[18]:

	content	category
0	The sporting industry has come a long way sin...	technology
1	Asian quake hits European shares Shares in Eu...	business
2	BT is offering customers free internet teleph...	technology
3	Barclays shares up on merger talk Shares in U...	business
4	England centre Olly Barkley has been passed f...	sport
...	...	...
1403	Woodward eyes Brennan for Lions Toulouse's fo...	sport
1404	The trial of Bernie Ebbers, former chief exec...	business
1405	Yukos accused of lying to court Russian oil f...	business
1406	Russian oil company Yukos has dropped the thr...	business
1407	Zambia's technical director, Kalusha Bwalya i...	sport

1408 rows × 2 columns

Tokenizing Text

In [19]:
```
tokenize = CountVectorizer().build_tokenizer()
stopwords = text.ENGLISH_STOP_WORDS
all_filtered_tokens = []
for doc in df['content']:
    tokens = tokenize(doc.lower())
    filtered_tokens = []
    for token in tokens:
        if not token in stopwords:
            filtered_tokens.append(token)
    all_filtered_tokens.append(' '.join(filtered_tokens))
print("Created %d filtered token lists" % len(all_filtered_tokens))
all_filtered_tokens
```

```
Created 1408 filtered token lists
```

Out[19]:
```
['sporting industry come long way 60s carved niche roots deep fathom sports industry showing sign decline time soon later reason seemingly subtle difference industries customers sporting industry fans vivek ranadivé leader ownership group nba sacramento kings explained beautifully fans paint face purple fans evangelize ceo business dying position dying fans fan passion certainly industry going leagues sporting franchises decided rest laurels years seen steady introduction technology world sports amplifying fans appreciation games enhancing athletes public profiles informing training methods influencing contests waged digital technology particular helped create alternative source revenue games corporate sponsorship achieved capitalizing ardor customer base sorry fan base',
 'asian quake hits european shares shares europe leading reinsurers travel firms fallen scale damage wrought tsunamis south asia apparent 23 000 people killed following massive underwater earthquake worst hit areas popular tourist destinations reisurance firms swiss munich lost value investors worried rebuilding costs disaster little impact stock markets asia currencies including thai baht indonesian rupiah weakened analysts warned economic growth slow came worst possible time said hans goetti singapore based fund manager impact tourist industry pretty devastating especially thailand travel related shares dropped europe companies germany tui lufthansa france club mediterranne sliding insurers reinsurance firms pressure europe shares munich swiss world biggest reinsurers fell market speculated cost rebuilding asia zurich financial allianz axa suffered decline value losses smaller reflecting market view reinsurers likely pick bulk costs worries size insurance liabilities dragged european shares impact exacerbated light post christmas trading germany benchmark dax index closed day 16 29 points lower 817 69 france cac index leading shares fell 07 points 817 69 investors pointed declines probably industry specific travel insurance firms hit hardest early concrete damage figures swiss spokesman floiran woest told associated press fact damage widely spread geogr
```

In [20]:
```
df['clean'] = all_filtered_tokens
df
```

Out[20]:

	content	category	clean
0	The sporting industry has come a long way sin...	technology	sporting industry come long way 60s carved nic...
1	Asian quake hits European shares Shares in Eu...	business	asian quake hits european shares shares europe...
2	BT is offering customers free internet teleph...	technology	bt offering customers free internet telephone ...
3	Barclays shares up on merger talk Shares in U...	business	barclays shares merger talk shares uk banking ...
4	England centre Olly Barkley has been passed f...	sport	england centre olly barkley passed fit sunday ...
...	...	...	...
1403	Woodward eyes Brennan for Lions Toulouse's fo...	sport	woodward eyes brennan lions toulouse irish int...
1404	The trial of Bernie Ebbers, former chief exec...	business	trial bernie ebbers chief executive bankrupt p...
1405	Yukos accused of lying to court Russian oil f...	business	yukos accused lying court russian oil firm yuk...
1406	Russian oil company Yukos has dropped the thr...	business	russian oil company yukos dropped threat legal...
1407	Zambia's technical director, Kalusha Bwalya i...	sport	zambia technical director kalusha bwalya confi...

1408 rows × 3 columns

Term Weighting

In [21]:
```
vectorizer = TfidfVectorizer(stop_words="english",min_df = 5)
X = vectorizer.fit_transform(df['clean'])
print(X)
```

```
  (0, 5659)     0.12635917755557236
  (0, 708)      0.20685338848350532
  (0, 1626)     0.10054293773849095
  (0, 275)      0.10054293773849095
  (0, 5721)     0.12035408877745242
  (0, 1533)     0.08746627050335351
  (0, 5139)     0.08974570153347976
  (0, 5668)     0.0905643964991388
  (0, 431)      0.10588530273148258
  (0, 1575)     0.07900257323550364
  (0, 2928)     0.07738720046898899
  (0, 4343)     0.08641357455116963
  (0, 1845)     0.0689235032011391
  (0, 3910)     0.11713634041489018
  (0, 6261)     0.08182421776046826
  (0, 4767)     0.07281830699778001
  (0, 595)      0.09629300380643811
  (0, 2157)     0.11569618005034049
  (0, 2668)     0.11616721188662793
  (0, 6715)     0.04208676868346757
```

Apprying Model

split data เพื่อนำไปทำ training และ testing data เพื่อไปเข้าโมเดล

In [22]:
```
Y = df['category']
```

In [23]:
```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=0)
```

Decision Tree

In [24]:
```
param_grid ={
    'criterion':['gini', 'entropy'],
    'max_depth':[2, 4, 6, 8, 10, 12]
}

model = DecisionTreeClassifier()
grid = GridSearchCV(model, param_grid, n_jobs=-1)
grid.fit(X_train, y_train)
print(grid.best_params_)
print(grid.best_score_)
```

```
{'criterion': 'gini', 'max_depth': 12}
0.864117994100295
```

In [25]:
```
Dtree = DecisionTreeClassifier(criterion = 'gini', max_depth=12)
Dtree.fit(X_train, y_train)
X_predict = Dtree.predict(X_test)

Dtree_accuracy = accuracy_score(X_predict, y_test)
Dtree_precision = precision_score(X_predict, y_test, average = 'micro')
Dtree_recall = recall_score(X_predict, y_test, average = 'micro')
```

In [26]:
```
print(classification_report(y_test, X_predict))
```

```
              precision    recall  f1-score   support

    business       0.79      0.81      0.80        95
       sport       0.92      0.89      0.91       108
  technology       0.81      0.82      0.82        79

    accuracy                           0.84       282
   macro avg       0.84      0.84      0.84       282
weighted avg       0.85      0.84      0.84       282
```

K-Nearest Neighbors (KNN)

In [27]:
```
param_grid ={
    'n_neighbors':[3, 5, 7]
}
model = KNeighborsClassifier()
grid = GridSearchCV(model, param_grid, n_jobs=-1)
grid.fit(X_train, y_train)
print(grid.best_params_)
print(grid.best_score_)
```

```
{'n_neighbors': 7}
0.9689085545722713
```

In [28]:
```
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
X_predict = knn.predict(X_test)

knn_accuracy = accuracy_score(X_predict, y_test)
knn_precision = precision_score(X_predict, y_test, average = 'micro')
knn_recall = recall_score(X_predict, y_test, average = 'micro')
```

In [29]:
```
print(classification_report(y_test, X_predict))
```

```
              precision    recall  f1-score   support

    business       0.96      0.93      0.94        95
       sport       0.96      0.95      0.96       108
  technology       0.92      0.96      0.94        79

    accuracy                           0.95       282
   macro avg       0.94      0.95      0.95       282
weighted avg       0.95      0.95      0.95       282
```

Random Forest

In [30]:
```
param_grid ={
    'max_depth':[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None],
    'max_features':['auto', 'sqrt'],
    'min_samples_leaf': [1, 2, 4]
}
model = RandomForestClassifier()
grid = GridSearchCV(model, param_grid, n_jobs=-1)
grid.fit(X_train, y_train)
print(grid.best_params_)
print(grid.best_score_)
```

```
{'max_depth': 70, 'max_features': 'sqrt', 'min_samples_leaf': 1}
0.9742536873156343
```

In [31]:
```
forest = RandomForestClassifier(max_depth=20, max_features='auto', min_samples_leaf=2)
forest.fit(X_train, y_train)
X_predict = forest.predict(X_test)

forest_accuracy = accuracy_score(X_predict, y_test)
forest_precision = precision_score(X_predict, y_test, average = 'micro')
forest_recall = recall_score(X_predict, y_test, average = 'micro')
```

In [32]:
```
print(classification_report(y_test, X_predict))
```

```
              precision    recall  f1-score   support

    business       0.96      0.97      0.96        95
       sport       0.95      0.98      0.97       108
  technology       0.96      0.91      0.94        79

    accuracy                           0.96       282
   macro avg       0.96      0.95      0.96       282
weighted avg       0.96      0.96      0.96       282
```

Logistic Regression

In [33]:
```
param_grid ={
    'penalty': ['l1', 'l2'],
    'max_iter': [100, 200, 300, 400, 500]
}
model = LogisticRegression(random_state=0)
grid = GridSearchCV(model, param_grid, n_jobs=-1)
grid.fit(X_train, y_train)
print(grid.best_params_)
print(grid.best_score_)
```

```
{'max_iter': 100, 'penalty': 'l2'}
0.9795791543756145
```

In [34]:
```
logis = LogisticRegression(max_iter=100, penalty='l2')
logis.fit(X_train, y_train)
X_predict = logis.predict(X_test)

logis_accuracy = accuracy_score(X_predict, y_test)
logis_precision = precision_score(X_predict, y_test, average = 'micro')
logis_recall = recall_score(X_predict, y_test, average = 'micro')
```

In [35]:
```
print(classification_report(y_test, X_predict))
```

```
              precision    recall  f1-score   support

    business       0.97      0.99      0.98        95
       sport       0.99      0.97      0.98       108
  technology       0.99      0.99      0.99        79

    accuracy                           0.98       282
   macro avg       0.98      0.98      0.98       282
weighted avg       0.98      0.98      0.98       282
```

In [ ]:
```
```