Context:

The Gurugram-based company 'FlipItNews' aims to revolutionize the way Indians perceive finance, business, and capital market investment, by giving it a boost through artificial intelligence (AI) and machine learning (ML). They're on a mission to reinvent financial literacy for Indians, where financial awareness is driven by smart information discovery and engagement with peers. Through their smart content discovery and contextual engagement, the company is simplifying business, finance, and investment for millennials and first-time investors

Objective:

The goal of this project is to use a bunch of news articles extracted from the companies' internal database and categorize them into several categories like politics, technology, sports, business and entertainment based on their content. Use natural language processing and create & compare at least three different models.

How NLP Enhances FlipItNews Operations and User Experience

1. Automated News Curation & Summarization

- **Problem:** Manually reading and summarizing financial news is slow and labor-intensive.
- **NLP Solution:**
  - **Text Summarization models** (extractive & abstractive) can condense long financial articles into 60–100 word summaries without losing key details.
  - Helps FlipItNews deliver "news in short" faster than competitors.

---

2. Real-Time Sentiment Analysis

- **Problem:** Investors want to know the market mood instantly.
- **NLP Solution:**
  - Sentiment analysis on stock-related news, press releases, and tweets to determine **positive, negative, or neutral** sentiment.
  - FlipItNews can show a **"Market Sentiment Index"** alongside news stories.

---

3. Topic Classification & Tagging

- **Problem:** Users prefer news tailored to their interests (e.g., Banking, IT, Pharma).
- **NLP Solution:**

    - Text classification models can automatically tag each article with sectors, companies, and keywords.
    - Improves **personalized recommendations** in the FlipItNews app.

---

4. **Fake News Detection**

- **Problem:** Financial misinformation can harm investors and the company's credibility.
- **NLP Solution:**
    - Use NLP + fact-checking databases to detect suspicious claims.
    - Flag or block unverified content before publication.

---

5. **Personalized News Feeds**

- **Problem:** Information overload can make users disengage.
- **NLP Solution:**
    - Build **user profiles** from reading history.
    - Recommend news articles based on their preferred sectors, companies, and sentiment preference.

---

6. **Speech-to-Text for Financial Briefings**

- **Problem:** Many financial updates come from company earnings calls or analyst briefings.
- **NLP Solution:**
    - Convert speech to text using

Attribute Information:

- Article
- Category

The features names are themselves pretty self-explanatory

Concepts Used:

- Natural Language Processing
- Text Processing
    - Stopwords, Tokenization, Lemmatization
    - Bag of Words, TF-IDF

- Multi-class Classification

Abstract:

- Installing & Importing all the required libraries and Loading the dataset.
- Conduct a preliminary analysis to understand the structure of the dataset and the distribution of news articles in each category.
- Create a user defined function to process the textual data (news articles).
    - Remove non-letters
    - Remove Stopwords
    - Word Tokenize the text
    - Perform Lemmatization
- Display how a single news article looks like before and after the processing.
- Encode the target variable (category) using Label/Ordinal encoder.
- Create an option for the user to choose between Bag of Words and TF-IDF techniques for vectorizing the data.
- Perform train-test split and train a Naive Bayes classifier model using the simple/classical approach.
- Evaluate the model's performance and plot the Confusion Matrix as well as Classification Report.
- Functionalize the code and train & evaluate three more classifier models (Decision Tree, Nearest Neighbors, Random Forest).
- Observe and comment on the performances of all the models used.

Installing the libraraies

In [225…

```
!pip install --user -U nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages
(3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages
(from nltk) (8.2.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages
(from nltk) (1.5.1)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-
packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (f
rom nltk) (4.67.1)
```

Importing all the requried libraries

In [226…

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
import os
import random
import string
import nltk
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
import plotly.express as px
from google.colab import drive
nltk.download('punkt')
from nltk.tokenize import sent_tokenize
from nltk.corpus import twitter_samples    # sample Twitter dataset from NLTK
from nltk.corpus import stopwords          # module for stop words that come wit
from nltk.stem import PorterStemmer        # module for stemming
from nltk.stem import WordNetLemmatizer    # module for Lemmatization
from nltk.tokenize import TweetTokenizer
nltk.download('stopwords')
nltk.download('wordnet')
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, roc_auc_score, f1_score, \
          precision_score, recall_score, classification_report, confusion_matrix

from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Mouting the Google Drive

In [227…

```
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call dri
ve.mount("/content/drive", force_remount=True).
```

In [228…
```
df = pd.read_csv('/content/drive/MyDrive/Scaler_DSML_Digital_Notes/BusinessCases
df.head()
```

Out[228…

	Category	Article
0	Technology	tv future in the hands of viewers with home th...
1	Business	worldcom boss left books alone former worldc...
2	Sports	tigers wary of farrell gamble leicester say ...
3	Sports	yeading face newcastle in fa cup premiership s...
4	Entertainment	ocean s twelve raids box office ocean s twelve...

Basic Analysis on the Given data

Shape

In [229…
```
df.shape
```

Out[229…
```
(2225, 2)
```

Info

In [230…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2225 entries, 0 to 2224
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Category  2225 non-null   object
 1   Article   2225 non-null   object
dtypes: object(2)
memory usage: 34.9+ KB
```

Type Caste

In [231…
```
df = df.astype({'Category':'category', 'Article':'string'})
df.head(5)
```

Out[231…

	Category	Article
0	Technology	tv future in the hands of viewers with home th...
1	Business	worldcom boss left books alone former worldc...
2	Sports	tigers wary of farrell gamble leicester say ...
3	Sports	yeading face newcastle in fa cup premiership s...
4	Entertainment	ocean s twelve raids box office ocean s twelve...

Observation From the above information we can see that there are no missing samples

Ditribution of Categories

In [232…

```
px.pie(df, names='Category',hole=0.5, title='Distribution of Categories')
```

Observation

- From the above we can understand that all classes are almost equal but Business and Sports classes are slightly higher

Encoding Categories

In [233…
```
cat_encode_dict = {'Technology':1, 'Business':2, 'Sports':3, 'Entertainment':4,
df['Encoded_Cat'] = df['Category'].map(cat_encode_dict)
df.head()
```

Out[233…

	Category	Article	Encoded_Cat
0	Technology	tv future in the hands of viewers with home th...	1
1	Business	worldcom boss left books alone former worldc...	2
2	Sports	tigers wary of farrell gamble leicester say ...	3
3	Sports	yeading face newcastle in fa cup premiership s...	3
4	Entertainment	ocean s twelve raids box office ocean s twelve...	4

Observation:

- All the classes are encoded properly
    - Technology - 1
    - Business - 2
    - Sports - 3
    - Entertainment - 4
    - Politics - 5

In [234…
```
df.dtypes
```

Out[234…

	0
Category	category
Article	string[python]
Encoded_Cat	category

dtype: object

Observation:

- Converted the 'Article' to string data types
- Converted the 'Category' to category data type and encoded with the numerical values for modelling.

Preprocessing the Articles using Manual vectorization technique

Following are handled in the text data

- Stop wor
- Stock Market ticker

- retweet text [RT]
- Hyperlinks
- "#" sign
- Punctuations

In [235…
```
def process_tweet(doc):
    lemmatizer = WordNetLemmatizer()
    stopwords_english = stopwords.words('english')
    # remove stock market tickers like $GE
    doc = re.sub(r'\$\w*', '', doc)
    # remove old style retweet text "RT"
    doc = re.sub(r'^RT[\s]+', '', doc)
    # remove hyperlinks
    doc = re.sub(r'https?://[^\s\n\r]+', '', doc)
    # remove hashtags
    # only removing the hash # sign from the word
    doc = re.sub(r'#', '', doc)
    # tokenize tweets
    tokenizer = TweetTokenizer(preserve_case=False, strip_handles=True,
                               reduce_len=True)
    doc_tokens = tokenizer.tokenize(doc)

    doc_clean = []
    for word in doc_tokens:
        if (word not in stopwords_english and  # remove stopwords
                word not in string.punctuation):  # remove punctuation
            # tweets_clean.append(word)
            lemma_word = lemmatizer.lemmatize(word)  # stemming word
            doc_clean.append(lemma_word)

    return doc_clean
```

In [236…
```
df['Article_Cleaned'] = df['Article'].apply(lambda x: process_tweet(x))
```

In [237…
```
df.head()
```

Out[237…

	Category	Article	Encoded_Cat	Article_Cleaned
0	Technology	tv future in the hands of viewers with home th...	1	[tv, future, hand, viewer, home, theatre, syst...
1	Business	worldcom boss left books alone former worldc...	2	[worldcom, bos, left, book, alone, former, wor...
2	Sports	tigers wary of farrell gamble leicester say ...	3	[tiger, wary, farrell, gamble, leicester, say,...
3	Sports	yeading face newcastle in fa cup premiership s...	3	[yeading, face, newcastle, fa, cup, premiershi...
4	Entertainment	ocean s twelve raids box office ocean s twelve...	4	[ocean, twelve, raid, box, office, ocean, twel...

Observation

Handled all the Articles properly by removing the below

- Stop words
- Stock Market ticker
- retweet text [RT]
- Hyperlinks
- "#" sign
- Punctuations

In [238…
```
processed_df = df[['Article_Cleaned', 'Encoded_Cat']]
processed_df.head()
```

Out[238…

	Article_Cleaned	Encoded_Cat
0	[tv, future, hand, viewer, home, theatre, syst...	1
1	[worldcom, bos, left, book, alone, former, wor...	2
2	[tiger, wary, farrell, gamble, leicester, say,...	3
3	[yeading, face, newcastle, fa, cup, premiershi...	3
4	[ocean, twelve, raid, box, office, ocean, twel...	4

In [239…
```
cat_encode_dict
```

Out[239…
```
{'Technology': 1,
 'Business': 2,
 'Sports': 3,
 'Entertainment': 4,
 'Politics': 5}
```

In [240…
```
technology_articles = processed_df[processed_df['Encoded_Cat'] == cat_encode_dic
business_articles = processed_df[processed_df['Encoded_Cat'] == cat_encode_dict[
sports_articles = processed_df[processed_df['Encoded_Cat'] == cat_encode_dict['S
entertainment_articles = processed_df[processed_df['Encoded_Cat'] == cat_encode_
politics_articles = processed_df[processed_df['Encoded_Cat'] == cat_encode_dict[
```

In [241…
```
for _ in [technology_articles, business_articles, sports_articles, entertainment
  display(_.head(2))
```

	Article_Cleaned	Encoded_Cat
0	[tv, future, hand, viewer, home, theatre, syst...	1
19	[game, maker, fight, survival, one, britain, l...	1

	Article_Cleaned	Encoded_Cat
1	[worldcom, bos, left, book, alone, former, wor...	2
11	[virgin, blue, share, plummet, 20, share, aust...	2

	Article_Cleaned	Encoded_Cat
2	[tiger, wary, farrell, gamble, leicester, say,...	3
3	[yeading, face, newcastle, fa, cup, premiershi...	3

	Article_Cleaned	Encoded_Cat
4	[ocean, twelve, raid, box, office, ocean, twel...	4
9	[last, star, war, child, sixth, final, star, w...	4

	Article_Cleaned	Encoded_Cat
5	[howard, hit, back, mongrel, jibe, michael, ho...	5
6	[blair, prepares, name, poll, date, tony, blai...	5

In [242…
```
all_cat_df_dict = {'Technology':technology_articles, 'Business':business_article

for _cat, _df in all_cat_df_dict.items():
  print(f'{_cat} Articles Shape : {_df.shape}')
```

```
Technology Articles Shape : (401, 2)
Business Articles Shape : (510, 2)
Sports Articles Shape : (511, 2)
Entertainment Articles Shape : (386, 2)
Politics Articles Shape : (417, 2)
```

Balancing the class imbalance

In [243…
```
#Balancing all the samples of different categories
bal_all_cat_dict = {_cat:list(_df['Article_Cleaned'][:386]) for _cat, _df in all

print('After Handling Imbalance data\n')
for _cat, _data_list in bal_all_cat_dict.items():
  print(f'{_cat} Articles Shape : {len(_data_list)}')

#The additional samples are used for testing
add_bal_all_cat_dict = {_cat:list(_df['Article_Cleaned'][386:]) for _cat, _df in
print('\nLength of the additional samples after balancing the data\n')
for _cat, _data_list in add_bal_all_cat_dict.items():
  print(f'{_cat} Articles Shape : {len(_data_list)}')
```

```
After Handling Imbalance data

Technology Articles Shape : 386
Business Articles Shape : 386
Sports Articles Shape : 386
Entertainment Articles Shape : 386
Politics Articles Shape : 386

Length of the additional samples after balancing the data

Technology Articles Shape : 15
Business Articles Shape : 124
Sports Articles Shape : 125
Entertainment Articles Shape : 0
Politics Articles Shape : 31
```

In [244…
```
# pd.Series(all_tech_articles, name='Article_Cleaned').head()

tech_df = pd.DataFrame({'Article_Cleaned' : bal_all_cat_dict['Technology'], 'Cat
business_df = pd.DataFrame({'Article_Cleaned' : bal_all_cat_dict['Business'], 'C
sports_df = pd.DataFrame({'Article_Cleaned' : bal_all_cat_dict['Sports'], 'Categ
entertainment_df = pd.DataFrame({'Article_Cleaned' : bal_all_cat_dict['Entertain
politics_df = pd.DataFrame({'Article_Cleaned' : bal_all_cat_dict['Politics'], 'C

tech_test_df1 = pd.DataFrame({'Article_Cleaned' : add_bal_all_cat_dict['Technolc
business_test_df1 = pd.DataFrame({'Article_Cleaned' : add_bal_all_cat_dict['Busi
sports_test_df1 = pd.DataFrame({'Article_Cleaned' : add_bal_all_cat_dict['Sports
politics_test_df1 = pd.DataFrame({'Article_Cleaned' : add_bal_all_cat_dict['Poli
```

In [245…
```
tech_df.head(2)
```

Out[245…

	Article_Cleaned	Category
0	[tv, future, hand, viewer, home, theatre, syst...	1
1	[game, maker, fight, survival, one, britain, l...	1

In [246…
```
business_df.head(2)
```

Out[246…

	Article_Cleaned	Category
0	[worldcom, bos, left, book, alone, former, wor...	2
1	[virgin, blue, share, plummet, 20, share, aust...	2

In [247…
```
sports_df.head(2)
```

Out[247…

	Article_Cleaned	Category
0	[tiger, wary, farrell, gamble, leicester, say,...	3
1	[yeading, face, newcastle, fa, cup, premiershi...	3

In [248…
```
entertainment_df.head(2)
```

Out[248...

	Article_Cleaned	Category
0	[ocean, twelve, raid, box, office, ocean, twel...	4
1	[last, star, war, child, sixth, final, star, w...	4

In [249...
```
politics_df.head(2)
```

Out[249...

	Article_Cleaned	Category
0	[howard, hit, back, mongrel, jibe, michael, ho...	5
1	[blair, prepares, name, poll, date, tony, blai...	5

In [250...
```
business_test_df1.head(2)
```

Out[250...

	Article_Cleaned	Category
0	[u, airway, staff, agree, pay, cut, union, rep...	2
1	[call, overhaul, uk, state, pension, uk, pensi...	2

In [251...
```
tech_train_df, tech_test_df2 = train_test_split(tech_df, test_size=0.2, random_s
business_train_df, business_test_df2 = train_test_split(business_df, test_size=0
sports_train_df, sports_test_df2 = train_test_split(sports_df, test_size=0.2, ra
entertainment_train_df, entertainment_test_df = train_test_split(entertainment_d
politics_train_df, politics_test_df2 = train_test_split(politics_df, test_size=0
```

In [252...
```
df_train = pd.concat([tech_train_df, business_train_df,  sports_train_df, entert

tech_test_df = pd.concat([tech_test_df1, tech_test_df2])
business_test_df = pd.concat([business_test_df1, business_test_df2])
sports_test_df = pd.concat([sports_test_df1, sports_test_df2])
politics_test_df = pd.concat([politics_test_df1, politics_test_df2])

df_test = pd.concat([tech_test_df, business_test_df,  sports_test_df, entertainm
```

In [253...
```
df_test
```

Out[253...

	Article_Cleaned	Category
0	[re-draft, eu, patent, law, proposed, european...	1
1	[game, win, blu-ray, dvd, format, next-generat...	1
2	[mobile, tv, tipped, one, watch, scandinavian,...	1
3	[lifestyle, governs, mobile, choice, faster, b...	1
4	[google, launch, tv, search, service, net, sea...	1
...	...	...
18	[uk, pledge, £, 1bn, vaccine, effort, uk, chan...	5
137	[child, access, law, shake-up, parent, refuse,...	5
357	[campbell, e-mail, row, silly, fuss, ex-no, 10...	5
168	[campaign, cold, call, questioned, labour, con...	5
63	[blunkett, hint, election, call, ex-home, secr...	5

685 rows × 2 columns

In [254...

```
df_train_shuffled = df_train.sample(frac=1, random_state=42).reset_index(drop=Tr
df_test_shuffled = df_test.sample(frac=1, random_state=42).reset_index(drop=True
df_train_shuffled.head(5)
```

Out[254...

	Article_Cleaned	Category
0	[viewer, able, shape, tv, imagine, editing, ti...	1
1	[indie, film, nomination, announced, mike, lei...	4
2	[net, regulation, still, possible, blurring, b...	1
3	[format, war, could, confuse, user, technology...	1
4	[russian, film, win, bbc, world, prize, russia...	4

In [255...

```
df_test_shuffled.head(5)
```

Out[255...

	Article_Cleaned	Category
0	[cairn, energy, indian, gas, find, share, cair...	2
1	[jarre, join, fairytale, celebration, french, ...	4
2	[christmas, shopper, flock, till, shop, uk, re...	2
3	[agassi, fear, melbourne, andre, agassi, invol...	3
4	[salary, scandal, cameroon, cameroon, say, wid...	2

In [256...

```
df_train_shuffled.shape
```

Out[256...

```
(1540, 2)
```

In [257…
```
df_test_shuffled.shape
```

Out[257…
```
(685, 2)
```

In [258…
```
px.pie(df_train_shuffled, names='Category',hole=0.5, title='Distribution of Trai
```

Observtion:

All the train data articles classes are balanced.

In [259…
```
px.pie(df_test_shuffled, names='Category',hole=0.5, title='Distribution of Test
```

Observation:

Additional samples after balancing the train data are added to the test data.

In [260…
```
X_train, y_train = df_train_shuffled['Article_Cleaned'],  df_train_shuffled['Cat
X_train.head(2)
```

Out[260…

	Article_Cleaned
0	[viewer, able, shape, tv, imagine, editing, ti...
1	[indie, film, nomination, announced, mike, lei...

dtype: object

In [261…
```
y_train.head(2)
```

Out[261…

	Category
0	1
1	4

dtype: int64

In [262…
```
X_test, y_test = df_test_shuffled['Article_Cleaned'],  df_test_shuffled['Categor
X_test.head(2)
```

Out[262…

	Article_Cleaned
0	[cairn, energy, indian, gas, find, share, cair...
1	[jarre, join, fairytale, celebration, french, ...

dtype: object

In [263…
```
y_test.head(2)
```

Out[263…

	Category
0	2
1	4

dtype: int64

In [264…
```
from collections import Counter

def get_word_freq_dict(X, y):
    words_list = []

    data = X.to_list()
    cat = y.to_list()

    for i in range(len(data)):
        for word in data[i]:
            words_list.append((word, cat[i]))

    word_pair_dict = Counter(words_list)

    print('Total Length of word pairs:', len(words_list))
    print('Unique Length of word pairs:', len(word_pair_dict))
    print(f'Length of word pair dictionary : {len(word_pair_dict)}')

    return dict(word_pair_dict)
```

In [265…
```
final_pair_freq_dict = get_word_freq_dict(X_train, y_train)
final_pair_freq_dict
```

```
Total Length of word pairs: 349261
Unique Length of word pairs: 46169
Length of word pair dictionary : 46169
```

Out[265…

```
{('viewer', 1): 34,
 ('able', 1): 113,
 ('shape', 1): 5,
 ('tv', 1): 200,
 ('imagine', 1): 2,
 ('editing', 1): 6,
 ('titanic', 1): 1,
 ('watch', 1): 45,
 ('favourite', 1): 18,
 ('bit', 1): 28,
 ('cutting', 1): 3,
 ('slushier', 1): 1,
 ('moment', 1): 28,
 ('star', 1): 27,
 ('war', 1): 33,
 ('leave', 1): 12,
 ('bare', 1): 1,
 ('bone', 1): 1,
 ('action-fest', 1): 1,
 ('manipulating', 1): 1,
 ('film', 1): 113,
 ('make', 1): 294,
 ('personalised', 1): 2,
 ('movie', 1): 86,
 ('beginning', 1): 14,
 ('ambitious', 1): 4,
 ('new', 1): 402,
 ('7.5', 1): 4,
 ('euro', 1): 9,
 ('£', 1): 125,
 ('5.1', 1): 1,
 ('project', 1): 91,
 ('funded', 1): 2,
 ('european', 1): 87,
 ('union', 1): 10,
 ('medium', 1): 158,
 ('millennium', 1): 1,
 ('nm2', 1): 4,
 ('endgame', 1): 1,
 ('development', 1): 54,
 ('completely', 1): 17,
 ('genre', 1): 10,
 ('allow', 1): 50,
 ('audience', 1): 39,
 ('create', 1): 52,
 ('world', 1): 212,
 ('based', 1): 57,
 ('specific', 1): 14,
 ('interest', 1): 37,
 ('taste', 1): 9,
 ('participate', 1): 4,
 ('storyline', 1): 4,
 ('manipulate', 1): 3,
 ('plot', 1): 6,
 ('even', 1): 124,
 ('set', 1): 125,
 ('prop', 1): 1,
 ('show', 1): 155,
 ('bt', 1): 106,
 ('one', 1): 388,
```

```
('13', 1): 12,
('partner', 1): 12,
('involved', 1): 31,
('contributing', 1): 1,
('software', 1): 285,
('originally', 1): 7,
('designed', 1): 44,
('spot', 1): 28,
('anomaly', 1): 4,
('cctv', 1): 2,
('picture', 1): 82,
('us', 1): 47,
('content', 1): 171,
('recognition', 1): 9,
('algorithm', 1): 6,
('three-year', 1): 4,
('work', 1): 185,
('seven', 1): 23,
('production', 1): 22,
('develops', 1): 4,
('tool', 1): 82,
('edit', 1): 4,
('need', 1): 120,
('experimental', 1): 3,
('television', 1): 40,
('driven', 1): 14,
('text', 1): 46,
('message', 1): 123,
('participant', 1): 2,
('selected', 1): 4,
('word', 1): 39,
('impact', 1): 20,
('character', 1): 33,
('drama', 1): 4,
('interact', 1): 19,
('developed', 1): 49,
('finland', 1): 5,
('shown', 1): 19,
('finnish', 1): 9,
('another', 1): 68,
('team', 1): 43,
('bbc', 1): 115,
('big', 1): 72,
('budget', 1): 2,
('mervyn', 1): 1,
('peake', 1): 1,
('gothic', 1): 1,
('fantasy', 1): 3,
('gormenghast', 1): 1,
('re-engineered', 1): 1,
('people', 1): 726,
('choose', 1): 18,
('variety', 1): 16,
('edited', 1): 1,
('version', 1): 143,
('allowing', 1): 19,
('u', 1): 286,
('access', 1): 119,
('material', 1): 32,
('prove', 1): 19,
```

```
('technology', 1): 476,
('principle', 1): 4,
('explained', 1): 15,
('dr', 1): 54,
('doug', 1): 4,
('williams', 1): 4,
('technical', 1): 38,
('manager', 1): 30,
('relatively', 1): 13,
('dumb', 1): 1,
('box', 1): 44,
('receives', 1): 4,
('signal', 1): 23,
('teaching', 1): 2,
('machine', 1): 134,
('look', 1): 102,
('like', 1): 235,
('lego', 1): 1,
('block', 1): 17,
('reassembled', 1): 1,
('perfect', 1): 8,
('sense', 1): 13,
('said', 1): 1272,
('interactive', 1): 32,
('gaming', 1): 106,
('limited', 1): 27,
('form', 1): 49,
('usually', 1): 18,
('mean', 1): 141,
('vote', 1): 4,
('hoping', 1): 13,
('occupy', 1): 1,
('space', 1): 36,
('in-between', 1): 1,
('added', 1): 95,
('co-ordinator', 1): 3,
('peter', 1): 6,
('stollenmayer', 1): 1,
('would', 1): 375,
('radically', 1): 5,
('alter', 1): 2,
('role', 1): 23,
('directly', 1): 18,
('influence', 1): 9,
('see', 1): 133,
('hear', 1): 15,
('according', 1): 146,
('personal', 1): 82,
('wish', 1): 7,
('user', 1): 359,
('longer', 1): 27,
('passive', 1): 1,
('become', 1): 83,
('active', 1): 7,
('engagers', 1): 1,
('also', 1): 420,
('important', 1): 57,
('sophisticated', 1): 22,
('enough', 1): 44,
('obey', 1): 1,
```

```
('complex', 1): 12,
('rule', 1): 37,
('cinematography', 1): 1,
('john', 1): 13,
('wyver', 1): 3,
('producer', 1): 4,
('illumination', 1): 1,
('matter', 1): 20,
('stringing', 1): 1,
('together', 1): 52,
('romantic', 1): 2,
('action', 1): 82,
('portion', 1): 5,
('mr', 1): 415,
('know', 1): 55,
('fit', 1): 19,
('visually', 1): 4,
('observing', 1): 1,
('time-honoured', 1): 1,
('go', 1): 116,
('term', 1): 37,
('story', 1): 33,
('aesthetically', 1): 1,
('pleasing', 1): 1,
('planning', 1): 18,
('entitled', 1): 3,
('golden', 1): 5,
('age', 1): 26,
('renaissance', 1): 1,
('art', 1): 24,
('so-called', 1): 34,
('area', 1): 52,
('poetry', 1): 2,
('music', 1): 322,
('architecture', 1): 9,
('range', 1): 52,
('news', 1): 139,
('documentary', 1): 2,
('comedy', 1): 2,
('indie', 4): 10,
('film', 4): 777,
('nomination', 4): 123,
('announced', 4): 29,
('mike', 4): 28,
('leigh', 4): 27,
('award-winning', 4): 10,
('abortion', 4): 5,
('drama', 4): 74,
('vera', 4): 34,
('drake', 4): 34,
('scooped', 4): 7,
('seven', 4): 21,
('year', 4): 491,
('british', 4): 167,
('independent', 4): 36,
('award', 4): 410,
('venice', 4): 7,
('winner', 4): 92,
('face', 4): 18,
('stiff', 4): 1,
```

```
('competition', 4): 25,
('shane', 4): 6,
('meadow', 4): 4,
('critically', 4): 7,
('acclaimed', 4): 9,
('dead', 4): 36,
('man', 4): 75,
('shoe', 4): 6,
('received', 4): 45,
('eight', 4): 31,
('also', 4): 329,
('running', 4): 15,
('clutch', 4): 3,
('summer', 4): 23,
('love', 4): 76,
('stalker', 4): 1,
('enduring', 4): 8,
('ceremony', 4): 87,
('london', 4): 89,
('30', 4): 19,
('november', 4): 14,
('chosen', 4): 16,
('jury', 4): 9,
('chaired', 4): 1,
('cold', 4): 5,
('mountain', 4): 2,
('director', 4): 190,
('anthony', 4): 9,
('minghella', 4): 1,
('including', 4): 133,
('actress', 4): 138,
('cate', 4): 10,
('blanchett', 4): 13,
('helena', 4): 1,
('bonham-carter', 4): 1,
('recognise', 4): 5,
('film-making', 4): 6,
('britain', 4): 45,
('established', 4): 9,
('ago', 4): 33,
('nominee', 4): 61,
('reflect', 4): 6,
('growing', 4): 9,
('strength', 4): 5,
('diversity', 4): 3,
('filmmaking', 4): 3,
('said', 4): 700,
('bifa', 4): 1,
('founder', 4): 5,
('elliot', 4): 5,
('grove', 4): 1,
('commenting', 4): 1,
('nominated', 4): 81,
('added', 4): 79,
('selection', 4): 3,
('committee', 4): 12,
('harder', 4): 2,
('time', 4): 167,
('ever', 4): 41,
('narrowing', 4): 1,
```

```
('field', 4): 14,
('joining', 4): 6,
('best', 4): 492,
('climbing', 4): 1,
('documentary', 4): 36,
('touching', 4): 5,
('void', 4): 3,
('zombie', 4): 1,
('comedy', 4): 112,
('shaun', 4): 6,
('geoffrey', 4): 3,
('rush', 4): 6,
('win', 4): 87,
('actor', 4): 205,
('role', 4): 121,
('peter', 4): 15,
('seller', 4): 9,
('recent', 4): 33,
('biopic', 4): 16,
('life', 4): 100,
('death', 4): 46,
('australian', 4): 8,
('star', 4): 260,
('daniel', 4): 7,
('craig', 4): 4,
('phil', 4): 12,
('davis', 4): 30,
('ian', 4): 12,
('hart', 4): 3,
('blind', 4): 4,
('fight', 4): 20,
('paddy', 4): 3,
('considine', 4): 3,
('supporting', 4): 35,
('rare', 4): 6,
('u', 4): 300,
('scarlett', 4): 3,
('johansson', 4): 3,
('among', 4): 63,
('contender', 4): 9,
('girl', 4): 22,
('pearl', 4): 1,
('earring', 4): 1,
('fellow', 4): 22,
('include', 4): 68,
('imelda', 4): 24,
('staunton', 4): 34,
('natalie', 4): 11,
('press', 4): 11,
('anne', 4): 2,
('reid', 4): 2,
('mother', 4): 19,
('eva', 4): 1,
('birthistle', 4): 1,
('ae', 4): 1,
('fond', 4): 2,
('kiss', 4): 1,
('...', 4): 35,
('kevin', 4): 14,
('mcdonald', 4): 6,
```

```
('former', 4): 56,
('douglas', 4): 13,
('hickox', 4): 1,
('directorial', 4): 2,
('debut', 4): 55,
('seasoned', 4): 1,
('film-makers', 4): 9,
('roger', 4): 8,
('michell', 4): 2,
('pavel', 4): 1,
('pavlikowsky', 4): 1,
('challenge', 4): 12,
('harry', 4): 34,
('potter', 4): 27,
('author', 4): 18,
('jk', 4): 7,
('rowling', 4): 7,
('receive', 4): 17,
('special', 4): 45,
('contribution', 4): 12,
('industry', 4): 76,
('net', 1): 262,
('regulation', 1): 9,
('still', 1): 100,
('possible', 1): 42,
('blurring', 1): 2,
('boundary', 1): 3,
('internet', 1): 197,
('raise', 1): 16,
('question', 1): 24,
('watchdog', 1): 12,
('ofcom', 1): 15,
('move', 1): 78,
('closer', 1): 10,
('year', 1): 391,
('tv-quality', 1): 2,
('video', 1): 197,
('online', 1): 201,
('becomes', 1): 9,
('norm', 1): 1,
('debate', 1): 11,
('westminster', 1): 1,
('industry', 1): 141,
('considered', 1): 21,
('option', 1): 14,
('lord', 1): 9,
('currie', 1): 3,
('chairman', 1): 13,
('super-regulator', 1): 1,
('told', 1): 102,
('panel', 1): 21,
('protecting', 1): 6,
('always', 1): 45,
('primary', 1): 4,
('concern', 1): 30,
('despite', 1): 45,
('remit', 1): 1,
('disquiet', 1): 1,
('increased', 1): 11,
('among', 1): 45,
```

```
('service', 1): 391,
('provider', 1): 31,
('speech', 1): 38,
('made', 1): 129,
('recent', 1): 47,
('month', 1): 162,
('hinted', 1): 3,
('might', 1): 57,
('organised', 1): 13,
('association', 1): 28,
('ispa', 1): 1,
('possibility', 1): 11,
('challenge', 1): 42,
('arise', 1): 2,
('truly', 1): 9,
('blur', 1): 7,
('balance', 1): 4,
('struck', 1): 6,
('consumer', 1): 205,
('assess', 1): 2,
('risk', 1): 24,
('adopting', 1): 4,
('currently', 1): 86,
('exist', 1): 10,
('regulate', 1): 5,
('self-regulation', 1): 1,
('practice', 1): 13,
('discussion', 1): 11,
('study', 1): 44,
('suggest', 1): 9,
('many', 1): 282,
('eight', 1): 21,
('million', 1): 247,
('household', 1): 17,
('uk', 1): 197,
('could', 1): 372,
('adopted', 1): 11,
('broadband', 1): 176,
('end', 1): 94,
('2005', 1): 109,
('open', 1): 70,
('door', 1): 12,
('delivered', 1): 7,
('company', 1): 272,
('streaming', 1): 10,
('web', 1): 141,
('already', 1): 117,
('entertainment', 1): 74,
('division', 1): 10,
('distribute', 1): 19,
('come', 1): 111,
('source', 1): 52,
('bskyb', 1): 3,
('itv', 1): 2,
('head', 1): 41,
('andrew', 1): 14,
('burke', 1): 7,
('spoke', 1): 1,
('creating', 1): 19,
('platform', 1): 14,
```

```
('risque', 1): 1,
('celebrity', 1): 9,
('chef', 1): 1,
('serving', 1): 1,
('expletive', 1): 1,
('hot', 1): 6,
('dinner', 1): 2,
('surely', 1): 3,
('push', 1): 18,
('limit', 1): 24,
('fact', 1): 36,
('requested', 1): 4,
('gone', 1): 10,
('length', 1): 2,
('download', 1): 79,
('maybe', 1): 7,
('entirely', 1): 11,
('free', 1): 94,
('long', 1): 50,
('claimed', 1): 16,
('responsibility', 1): 12,
('carry', 1): 27,
('server', 1): 42,
('since', 1): 68,
('law', 1): 67,
('commission', 1): 21,
('dubbed', 1): 13,
('mere', 1): 3,
('conduit', 1): 1,
('back', 1): 81,
('2002', 1): 11,
('defence', 1): 11,
('apply', 1): 7,
('actual', 1): 9,
('knowledge', 1): 14,
('illegal', 1): 32,
('failed', 1): 14,
('remove', 1): 17,
('level', 1): 29,
('tested', 1): 9,
('several', 1): 76,
('high-profile', 1): 10,
('legal', 1): 85,
('case', 1): 44,
('richard', 1): 9,
('ayers', 1): 2,
('portal', 1): 18,
('director', 1): 67,
('tiscali', 1): 1,
('little', 1): 48,
('point', 1): 59,
('trying', 1): 25,
('impossible', 1): 7,
('huge', 1): 60,
('change', 1): 59,
('afoot', 1): 1,
('predicted', 1): 36,
('offer', 1): 125,
('planned', 1): 5,
('player', 1): 243,
```

```
('give', 1): 75,
('surfer', 1): 12,
('chance', 1): 24,
('programme', 1): 72,
('eastenders', 1): 3,
('top', 1): 85,
('gear', 1): 13,
('mainstream', 1): 16,
('whole', 1): 23,
('vast', 1): 17,
('sum', 1): 5,
('money', 1): 63,
('maintaining', 1): 6,
('network', 1): 236,
('supply', 1): 14,
('quantity', 1): 4,
('data', 1): 185,
('herald', 1): 3,
('digital', 1): 281,
('licence', 1): 10,
('fee', 1): 16,
('inappropriate', 1): 4,
('obviously', 1): 9,
('pornography', 1): 7,
('viewed', 1): 4,
('child', 1): 35,
('continues', 1): 11,
('dominate', 1): 8,
('headline', 1): 7,
('remains', 1): 18,
('political', 1): 22,
('issue', 1): 56,
('mp', 1): 1,
('allan', 1): 2,
('liberal', 1): 1,
('democrat', 1): 2,
('spokesman', 1): 36,
('think', 1): 86,
('answer', 1): 13,
('lie', 1): 7,
('somewhere', 1): 3,
('cry', 1): 3,
('offline', 1): 9,
('instead', 1): 52,
('seeing', 1): 22,
('brought', 1): 20,
('future', 1): 86,
('bring', 1): 26,
('departed', 1): 1,
('agreed', 1): 18,
('reality', 1): 13,
('power', 1): 76,
('likely', 1): 87,
('reign', 1): 2,
('on-demand', 1): 2,
('pulled', 1): 2,
('rather', 1): 62,
('pushed', 1): 10,
('choice', 1): 22,
('watershed', 1): 1,
```

```
('format', 1): 63,
('confuse', 1): 4,
('firm', 1): 325,
('sony', 1): 129,
('philip', 1): 8,
('matsushita', 1): 4,
('samsung', 1): 18,
('developing', 1): 31,
('common', 1): 29,
('way', 1): 268,
('stop', 1): 52,
('pirating', 1): 2,
('want', 1): 147,
('system', 1): 279,
('ensures', 1): 6,
('file', 1): 125,
('play', 1): 101,
('hardware', 1): 23,
('thwart', 1): 3,
('copying', 1): 8,
('confusion', 1): 4,
('faced', 1): 12,
('different', 1): 110,
('conflicting', 1): 2,
('control', 1): 104,
('expert', 1): 49,
('warned', 1): 29,
('say', 1): 232,
('guarantee', 1): 7,
('prevent', 1): 14,
('piracy', 1): 30,
('store', 1): 53,
('wrap', 1): 2,
('downloadable', 1): 9,
('own-brand', 1): 3,
('played', 1): 17,
('small', 1): 66,
('number', 1): 208,
('known', 1): 50,
('right', 1): 86,
('management', 1): 17,
('setting', 1): 26,
('alliance', 1): 12,
('hope', 1): 36,
('current', 1): 56,
('fragmentation', 1): 2,
('joint', 1): 12,
('statement', 1): 36,
('wanted', 1): 28,
('let', 1): 109,
('enjoy', 1): 13,
('appropriately', 1): 2,
('licensed', 1): 9,
('device', 1): 219,
('independent', 1): 16,
('obtained', 1): 4,
('harder', 1): 8,
('copy', 1): 44,
('bought', 1): 31,
('called', 1): 78,
```

```
('marlin', 1): 6,
('define', 1): 3,
('basic', 1): 18,
('specification', 1): 11,
('every', 1): 100,
('electronics', 1): 63,
('conform', 1): 4,
('built', 1): 13,
('intertrust', 1): 2,
('well', 1): 118,
('earlier', 1): 39,
('drm', 1): 18,
('group', 1): 99,
('coral', 1): 2,
('consortium', 1): 8,
('widely', 1): 36,
('seen', 1): 53,
('four', 1): 51,
('decide', 1): 7,
('destiny', 1): 2,
('sign', 1): 33,
('apple', 1): 149,
('microsoft', 1): 213,
('confusingly', 1): 2,
('sit', 1): 9,
('alongside', 1): 16,
('rival', 1): 37,
('akin', 1): 5,
('physical', 1): 11,
('betamax', 1): 10,
('vhs', 1): 6,
('past', 1): 39,
('ian', 1): 11,
('fogg', 1): 15,
('analyst', 1): 101,
('jupiter', 1): 23,
('research', 1): 142,
('difference', 1): 22,
('fragmented', 1): 2,
('two-horse', 1): 2,
('race', 1): 12,
('five', 1): 58,
('six', 1): 41,
('eight-horse', 1): 2,
('careful', 1): 7,
('buying', 1): 23,
('ensure', 1): 34,
('incompatibility', 1): 3,
('within', 1): 35,
('family', 1): 25,
('although', 1): 80,
('initiative', 1): 21,
('sure', 1): 31,
('program', 1): 158,
('help', 1): 143,
('uncertainty', 1): 3,
('life', 1): 87,
('confusing', 1): 8,
('time', 1): 249,
('shelley', 1): 2,
```

```
('taylor', 1): 5,
('author', 1): 10,
('report', 1): 118,
('lock', 1): 3,
('done', 1): 45,
('maximise', 1): 2,
('cash', 1): 66,
('itunes', 1): 30,
('example', 1): 34,
('hugely', 1): 23,
('successful', 1): 19,
('justify', 1): 4,
('existence', 1): 15,
('sell', 1): 29,
('ipod', 1): 81,
('rampant', 1): 4,
('competition', 1): 45,
('230', 1): 3,
('figure', 1): 60,
('drive', 1): 77,
('openness', 1): 2,
('freer', 1): 2,
('win', 1): 22,
('run', 1): 58,
('listen', 1): 26,
('earliest', 1): 2,
('m', 1): 57,
('place', 1): 62,
('explain', 1): 3,
('continuing', 1): 15,
('popularity', 1): 25,
('file-sharing', 1): 39,
('get', 1): 259,
('hold', 1): 53,
('pirated', 1): 25,
('pop', 1): 5,
('portability', 1): 7,
('peer-to-peer', 1): 23,
('100', 1): 43,
('cory', 1): 3,
('doctorow', 1): 6,
('electronic', 1): 37,
('frontier', 1): 6,
('foundation', 1): 17,
('campaign', 1): 55,
('cyber-rights', 1): 2,
('expressed', 1): 7,
('doubt', 1): 19,
('achieve', 1): 7,
('aim', 1): 34,
('ever', 1): 42,
('prevented', 1): 5,
('readily', 1): 2,
('admit', 1): 3,
('protection', 1): 25,
('skilled', 1): 3,
('attacker', 1): 4,
('crime', 1): 22,
('gang', 1): 9,
('responsible', 1): 11,
```

```
('intended', 1): 10,
('studio', 1): 51,
('label', 1): 8,
('perceive', 1): 5,
('opportunity', 1): 29,
('auto', 1): 10,
('american', 1): 40,
('ringtone', 1): 8,
('russian', 4): 8,
('bbc', 4): 149,
('world', 4): 105,
('prize', 4): 103,
('return', 4): 41,
('vozvrashchenie', 4): 2,
('named', 4): 58,
('four', 4): 78,
('cinema', 4): 45,
('tell', 4): 17,
('story', 4): 56,
('two', 4): 171,
('adolescent', 4): 2,
('boy', 4): 62,
('subjected', 4): 2,
('harsh', 4): 2,
('regime', 4): 5,
('strict', 4): 3,
('father', 4): 21,
('10', 4): 77,
('absence', 4): 5,
('directed', 4): 41,
('andrey', 4): 2,
('zvyagintsev', 4): 2,
('previously', 4): 27,
('2003', 4): 54,
('golden', 4): 47,
('lion', 4): 5,
('festival', 4): 114,
('presented', 4): 19,
('held', 4): 46,
('thursday', 4): 24,
('hosted', 4): 12,
('jonathan', 4): 11,
('ross', 4): 24,
('panel', 4): 16,
('included', 4): 51,
('x', 4): 10,
('file', 4): 3,
('gillian', 4): 2,
('anderson', 4): 3,
('critic', 4): 45,
('clarke', 4): 3,
('presenter', 4): 31,
('one', 4): 289,
('2005', 4): 32,
('involved', 4): 25,
('deliberation', 4): 2,
('shortlist', 4): 16,
('six', 4): 37,
('around', 4): 54,
('drawn', 4): 9,
```

```
('chose', 4): 7,
('motorcycle', 4): 10,
('diary', 4): 20,
('zatoichi', 4): 2,
('hero', 4): 21,
('viewer', 4): 44,
('poll', 4): 17,
('saw', 4): 44,
('zhang', 4): 7,
('yimou', 4): 3,
('martial', 4): 2,
('art', 4): 34,
('epic', 4): 14,
('emerge', 4): 4,
('favourite', 4): 45,
('32', 4): 2,
('vote', 4): 33,
('cast', 4): 26,
('tragedy', 4): 6,
('struck', 4): 4,
('production', 4): 56,
('young', 4): 63,
('15', 4): 15,
('year-old', 4): 52,
('vladimir', 4): 2,
('girin', 4): 2,
('drowned', 4): 3,
('lake', 4): 2,
('last', 4): 211,
('french', 4): 36,
('animated', 4): 16,
('feature', 4): 40,
('belleville', 4): 2,
('rendezvous', 4): 2,
('rapper', 4): 30,
('50', 4): 55,
('cent', 4): 31,
('end', 4): 52,
('protege', 4): 6,
('feud', 4): 8,
('ended', 4): 12,
('public', 4): 58,
('game', 4): 40,
('pair', 4): 21,
('wanted', 4): 30,
('good', 4): 70,
('model', 4): 8,
('community', 4): 17,
('row', 4): 11,
('blew', 4): 2,
('threw', 4): 3,
('g-unit', 4): 5,
('crew', 4): 7,
('accused', 4): 10,
('disloyal', 4): 2,
('member', 4): 49,
('entourage', 4): 4,
('reportedly', 4): 6,
('shot', 4): 29,
('outside', 4): 32,
```

```
('radio', 4): 74,
('station', 4): 36,
('interviewed', 4): 2,
('shook', 4): 3,
('hand', 4): 11,
('handed', 4): 14,
('money', 4): 61,
('music', 4): 378,
('project', 4): 33,
('new', 4): 238,
('york', 4): 47,
('deprived', 4): 2,
('area', 4): 12,
('wednesday', 4): 24,
('whose', 4): 31,
('real', 4): 36,
('name', 4): 52,
('jayceon', 4): 1,
('taylor', 4): 12,
('told', 4): 80,
('news', 4): 42,
('conference', 4): 6,
('want', 4): 49,
('apologise', 4): 3,
('almost', 4): 21,
('ashamed', 4): 2,
('participated', 4): 1,
('thing', 4): 45,
('went', 4): 59,
('week', 4): 102,
('chart-topper', 4): 5,
('curtis', 4): 11,
('jackson', 4): 45,
('truce', 4): 2,
('came', 4): 57,
('anniversary', 4): 15,
('notorious', 4): 1,
('big', 4): 68,
('1997', 4): 9,
('part', 4): 62,
('volatile', 4): 1,
('east', 4): 7,
('west', 4): 48,
('coast', 4): 2,
('rap', 4): 22,
('scene', 4): 38,
('today', 4): 19,
('show', 4): 323,
('people', 4): 181,
('rise', 4): 14,
('difficult', 4): 10,
('circumstance', 4): 1,
('together', 4): 17,
('put', 4): 32,
('negativity', 4): 1,
('behind', 4): 32,
('lot', 4): 32,
('see', 4): 59,
('happen', 4): 8,
('responding', 4): 1,
```

```
 ('important', 4): 22,
 ('group', 4): 76,
 ('family', 4): 61,
 ('fan', 4): 78,
 ('choir', 4): 5,
 ('harlem', 4): 5,
 ('got', 4): 49,
 ('cheque', 4): 2,
 ('000', 4): 106,
 ('£', 4): 198,
 ('77', 4): 2,
 ('800', 4): 3,
 ('500', 4): 7,
 ('53', 4): 3,
 ('400', 4): 5,
 ('made', 4): 119,
 ('compton', 4): 1,
 ('school', 4): 55,
 ('programme', 4): 54,
 ('launched', 4): 10,
 ('g-unity', 4): 1,
 ('foundation', 4): 7,
 ('help', 4): 28,
 ('overcome', 4): 3,
 ('obstacle', 4): 1,
 ('make', 4): 93,
 ('chance', 4): 18,
 ('better', 4): 24,
 ('realised', 4): 8,
 ('going', 4): 66,
 ('effective', 4): 2,
 ('need', 4): 26,
 ('set', 4): 71,
 ('example', 4): 10,
 ('stranger', 4): 2,
 ('ja', 4): 4,
 ('rule', 4): 11,
 ('target', 4): 4,
 ('ridicule', 4): 1,
 ('song', 4): 206,
 ...}
```

Observation:

All the frequencies pairs (word , class) are calulated and stored in final_pair_freq_dict

In [266…

```
words = []

for key in final_pair_freq_dict.keys():
  words.append(key[0])

unique_words = set(words)
print('Total Unique Words', len(unique_words))
```

```
Total Unique Words 25064
```

Observation:

- There are in total 25064 unique pair of (word,class).

In [267…

```
categories = ['Technology', 'Business', 'Sports', 'Entertainment', 'Politics']
word_cat_freq_list = []

for word in words:
    row = [word]
    for category in categories:
        freq = final_pair_freq_dict.get((word, cat_encode_dict[category]), 0)
        row.append(freq)
    word_cat_freq_list.append(row)
```

In [268…

```
word_cat_freq_list
```

Out[268…

```
[['viewer', 34, 0, 0, 44, 0],
 ['able', 113, 22, 31, 20, 53],
 ['shape', 5, 2, 8, 2, 5],
 ['tv', 200, 8, 8, 175, 5],
 ['imagine', 2, 0, 3, 3, 3],
 ['editing', 6, 0, 0, 5, 0],
 ['titanic', 1, 0, 0, 5, 1],
 ['watch', 45, 1, 11, 9, 8],
 ['favourite', 18, 1, 20, 45, 2],
 ['bit', 28, 3, 44, 14, 6],
 ['cutting', 3, 14, 3, 1, 14],
 ['slushier', 1, 0, 0, 0, 0],
 ['moment', 28, 3, 36, 24, 18],
 ['star', 27, 5, 38, 260, 4],
 ['war', 33, 7, 6, 43, 76],
 ['leave', 12, 11, 12, 15, 31],
 ['bare', 1, 0, 0, 0, 0],
 ['bone', 1, 1, 3, 6, 2],
 ['action-fest', 1, 0, 0, 0, 0],
 ['manipulating', 1, 1, 0, 0, 1],
 ['film', 113, 3, 1, 777, 4],
 ['make', 294, 78, 119, 93, 169],
 ['personalised', 2, 0, 0, 1, 2],
 ['movie', 86, 0, 0, 125, 0],
 ['beginning', 14, 11, 10, 8, 8],
 ['ambitious', 4, 2, 3, 3, 2],
 ['new', 402, 253, 151, 238, 328],
 ['7.5', 4, 5, 0, 0, 0],
 ['euro', 9, 126, 5, 9, 8],
 ['£', 125, 289, 69, 198, 286],
 ['5.1', 1, 4, 0, 1, 0],
 ['project', 91, 28, 0, 33, 20],
 ['funded', 2, 5, 1, 4, 5],
 ['european', 87, 109, 71, 25, 84],
 ['union', 10, 49, 31, 6, 87],
 ['medium', 158, 26, 14, 27, 34],
 ['millennium', 1, 0, 4, 4, 2],
 ['nm2', 4, 0, 0, 0, 0],
 ['endgame', 1, 0, 0, 1, 0],
 ['development', 54, 51, 3, 5, 33],
 ['completely', 17, 2, 8, 6, 14],
 ['genre', 10, 0, 0, 26, 0],
 ['allow', 50, 21, 15, 12, 62],
 ['audience', 39, 2, 0, 63, 11],
 ['create', 52, 31, 1, 9, 21],
 ['world', 212, 175, 223, 105, 96],
 ['based', 57, 20, 5, 32, 29],
 ['specific', 14, 5, 2, 6, 18],
 ['interest', 37, 109, 26, 18, 48],
 ['taste', 9, 0, 0, 4, 1],
 ['participate', 4, 1, 1, 0, 2],
 ['storyline', 4, 0, 0, 2, 0],
 ['manipulate', 3, 1, 0, 0, 1],
 ['plot', 6, 0, 0, 5, 3],
 ['even', 124, 49, 47, 43, 77],
 ['set', 125, 72, 137, 71, 99],
 ['prop', 1, 1, 14, 1, 0],
 ['show', 155, 32, 12, 323, 59],
 ['bt', 106, 15, 0, 0, 1],
 ['one', 388, 149, 206, 289, 229],
```

```
['13', 12, 16, 19, 28, 8],
['partner', 12, 8, 9, 13, 12],
['involved', 31, 18, 19, 25, 28],
['contributing', 1, 2, 0, 1, 1],
['software', 285, 7, 0, 0, 0],
['originally', 7, 6, 2, 12, 6],
['designed', 44, 14, 1, 8, 15],
['spot', 28, 1, 15, 15, 4],
['anomaly', 4, 0, 0, 0, 0],
['cctv', 2, 0, 0, 0, 0],
['picture', 82, 10, 2, 34, 7],
['us', 47, 1, 1, 4, 3],
['content', 171, 1, 5, 4, 7],
['recognition', 9, 1, 2, 5, 7],
['algorithm', 6, 0, 0, 0, 0],
['three-year', 4, 2, 0, 1, 0],
['work', 185, 52, 48, 79, 143],
['seven', 23, 9, 34, 21, 18],
['production', 22, 71, 0, 56, 0],
['develops', 4, 0, 0, 0, 0],
['tool', 82, 3, 0, 1, 2],
['edit', 4, 0, 0, 0, 0],
['need', 120, 72, 53, 26, 122],
['experimental', 3, 0, 0, 1, 0],
['television', 40, 7, 16, 58, 8],
['driven', 14, 16, 6, 5, 11],
['text', 46, 1, 0, 6, 7],
['message', 123, 6, 3, 9, 31],
['participant', 2, 1, 1, 0, 0],
['selected', 4, 1, 5, 0, 8],
['word', 39, 5, 9, 18, 19],
['impact', 20, 41, 8, 7, 17],
['character', 33, 2, 7, 23, 10],
['drama', 4, 0, 3, 74, 0],
['interact', 19, 0, 0, 0, 0],
['developed', 49, 7, 0, 5, 6],
['finland', 5, 1, 1, 1, 0],
['shown', 19, 13, 7, 24, 23],
['finnish', 9, 0, 0, 2, 0],
['another', 68, 39, 71, 38, 48],
['team', 43, 15, 193, 13, 15],
['bbc', 115, 37, 68, 149, 190],
['big', 72, 30, 54, 68, 48],
['budget', 2, 93, 1, 8, 69],
['mervyn', 1, 2, 1, 0, 0],
['peake', 1, 0, 0, 0, 0],
['gothic', 1, 0, 0, 1, 0],
['fantasy', 3, 1, 0, 6, 3],
['gormenghast', 1, 0, 0, 0, 0],
['re-engineered', 1, 0, 0, 0, 0],
['people', 726, 118, 58, 181, 451],
['choose', 18, 5, 3, 6, 7],
['variety', 16, 2, 3, 18, 2],
['edited', 1, 0, 0, 0, 2],
['version', 143, 1, 0, 53, 3],
['allowing', 19, 8, 3, 2, 12],
['u', 286, 501, 132, 300, 130],
['access', 119, 20, 0, 8, 30],
['material', 32, 10, 0, 20, 13],
['prove', 19, 4, 9, 3, 6],
```

```
['technology', 476, 12, 7, 3, 9],
['principle', 4, 4, 2, 0, 19],
['explained', 15, 9, 10, 6, 3],
['dr', 54, 11, 1, 10, 18],
['doug', 4, 0, 1, 0, 6],
['williams', 4, 2, 106, 38, 6],
['technical', 38, 7, 2, 8, 1],
['manager', 30, 18, 75, 13, 5],
['relatively', 13, 8, 2, 1, 1],
['dumb', 1, 0, 0, 3, 0],
['box', 44, 1, 14, 84, 2],
['receives', 4, 1, 1, 0, 1],
['signal', 23, 11, 2, 4, 8],
['teaching', 2, 1, 0, 2, 5],
['machine', 134, 2, 0, 2, 3],
['look', 102, 40, 51, 36, 56],
['like', 235, 34, 117, 114, 119],
['lego', 1, 0, 0, 0, 0],
['block', 17, 8, 6, 1, 5],
['reassembled', 1, 0, 0, 0, 0],
['perfect', 8, 1, 5, 4, 2],
['sense', 13, 7, 3, 7, 18],
['said', 1272, 1049, 590, 700, 1631],
['interactive', 32, 1, 0, 10, 1],
['gaming', 106, 0, 0, 2, 1],
['limited', 27, 20, 2, 5, 14],
['form', 49, 8, 46, 14, 40],
['usually', 18, 5, 3, 7, 5],
['mean', 141, 46, 22, 25, 71],
['vote', 4, 5, 3, 33, 130],
['hoping', 13, 4, 21, 16, 9],
['occupy', 1, 1, 0, 0, 1],
['space', 36, 11, 10, 2, 2],
['in-between', 1, 0, 0, 0, 0],
['added', 95, 80, 125, 79, 123],
['co-ordinator', 3, 0, 0, 1, 3],
['peter', 6, 15, 12, 15, 19],
['stollenmayer', 1, 0, 0, 0, 0],
['would', 375, 261, 234, 162, 763],
['radically', 5, 0, 0, 1, 1],
['alter', 2, 0, 1, 0, 0],
['role', 23, 21, 12, 121, 72],
['directly', 18, 6, 1, 2, 10],
['influence', 9, 3, 3, 15, 13],
['see', 133, 56, 76, 59, 92],
['hear', 15, 3, 2, 7, 7],
['according', 146, 73, 7, 39, 23],
['personal', 82, 13, 26, 14, 38],
['wish', 7, 8, 8, 3, 4],
['user', 359, 4, 0, 0, 1],
['longer', 27, 18, 4, 10, 26],
['passive', 1, 1, 0, 0, 1],
['become', 83, 28, 13, 59, 35],
['active', 7, 3, 0, 2, 1],
['engagers', 1, 0, 0, 0, 0],
['also', 420, 286, 208, 329, 330],
['important', 57, 23, 34, 22, 60],
['sophisticated', 22, 1, 0, 0, 0],
['enough', 44, 33, 36, 13, 40],
['obey', 1, 1, 0, 0, 1],
```

```
['complex', 12, 5, 1, 2, 10],
['rule', 37, 45, 13, 11, 91],
['cinematography', 1, 0, 0, 2, 0],
['john', 13, 18, 28, 50, 64],
['wyver', 3, 0, 0, 0, 0],
['producer', 4, 29, 0, 94, 1],
['illumination', 1, 0, 0, 0, 0],
['matter', 20, 10, 26, 4, 45],
['stringing', 1, 0, 0, 0, 0],
['together', 52, 12, 16, 17, 22],
['romantic', 2, 0, 0, 12, 0],
['action', 82, 46, 45, 36, 68],
['portion', 5, 0, 0, 3, 1],
['mr', 415, 384, 8, 162, 1291],
['know', 55, 14, 106, 51, 57],
['fit', 19, 3, 29, 6, 6],
['visually', 4, 0, 0, 0, 0],
['observing', 1, 0, 0, 0, 0],
['time-honoured', 1, 0, 0, 0, 1],
['go', 116, 42, 126, 79, 98],
['term', 37, 45, 14, 27, 58],
['story', 33, 4, 10, 56, 14],
['aesthetically', 1, 0, 0, 1, 0],
['pleasing', 1, 0, 1, 0, 0],
['planning', 18, 15, 2, 2, 31],
['entitled', 3, 4, 0, 6, 12],
['golden', 5, 12, 3, 47, 8],
['age', 26, 7, 5, 39, 36],
['renaissance', 1, 2, 0, 0, 0],
['art', 24, 2, 0, 34, 13],
['so-called', 34, 14, 1, 2, 5],
['area', 52, 45, 13, 12, 53],
['poetry', 2, 0, 0, 3, 0],
['music', 322, 3, 0, 378, 1],
['architecture', 9, 0, 0, 0, 0],
['range', 52, 14, 8, 12, 10],
['news', 139, 98, 20, 42, 82],
['documentary', 2, 0, 0, 36, 7],
['comedy', 2, 0, 0, 112, 0],
['indie', 0, 0, 0, 10, 0],
['film', 113, 3, 1, 777, 4],
['nomination', 3, 2, 1, 123, 0],
['announced', 46, 43, 6, 29, 59],
['mike', 12, 4, 23, 28, 9],
['leigh', 0, 0, 0, 27, 4],
['award-winning', 0, 1, 0, 10, 0],
['abortion', 0, 0, 0, 5, 10],
['drama', 4, 0, 3, 74, 0],
['vera', 0, 0, 2, 34, 0],
['drake', 0, 0, 0, 34, 0],
['scooped', 0, 0, 2, 7, 0],
['seven', 23, 9, 34, 21, 18],
['year', 391, 553, 245, 491, 329],
['british', 26, 25, 49, 167, 137],
['independent', 16, 16, 9, 36, 30],
['award', 34, 1, 15, 410, 4],
['venice', 0, 0, 0, 7, 1],
['winner', 8, 2, 33, 92, 6],
['face', 37, 44, 82, 18, 48],
['stiff', 3, 1, 1, 1, 2],
```

['competition', 45, 26, 31, 25, 5],
['shane', 0, 0, 17, 6, 0],
['meadow', 0, 0, 2, 4, 0],
['critically', 0, 0, 0, 7, 0],
['acclaimed', 4, 0, 0, 9, 0],
['dead', 6, 2, 3, 36, 5],
['man', 17, 14, 50, 75, 30],
['shoe', 1, 1, 0, 6, 1],
['received', 21, 20, 14, 45, 20],
['eight', 21, 9, 35, 31, 20],
['also', 420, 286, 208, 329, 330],
['running', 50, 19, 21, 15, 19],
['clutch', 3, 1, 1, 3, 0],
['summer', 22, 13, 59, 23, 7],
['love', 16, 0, 15, 76, 9],
['stalker', 1, 0, 0, 1, 0],
['enduring', 0, 0, 0, 8, 0],
['ceremony', 3, 2, 5, 87, 5],
['london', 41, 69, 25, 89, 78],
['30', 38, 45, 17, 19, 23],
['november', 41, 54, 19, 14, 25],
['chosen', 12, 5, 3, 16, 12],
['jury', 1, 2, 1, 9, 3],
['chaired', 0, 0, 0, 1, 5],
['cold', 2, 8, 3, 5, 12],
['mountain', 2, 3, 0, 2, 1],
['director', 67, 51, 32, 190, 29],
['anthony', 0, 6, 1, 9, 3],
['minghella', 0, 0, 0, 1, 0],
['including', 57, 34, 19, 133, 42],
['actress', 0, 0, 0, 138, 0],
['cate', 0, 0, 2, 10, 0],
['blanchett', 0, 0, 0, 13, 0],
['helena', 0, 1, 0, 1, 0],
['bonham-carter', 0, 0, 0, 1, 0],
['recognise', 11, 3, 5, 5, 13],
['film-making', 0, 0, 0, 6, 0],
['britain', 27, 5, 46, 45, 201],
['established', 9, 2, 3, 9, 5],
['ago', 34, 35, 33, 33, 25],
['nominee', 1, 0, 0, 61, 0],
['reflect', 1, 7, 0, 6, 8],
['growing', 57, 30, 1, 9, 11],
['strength', 2, 18, 5, 5, 6],
['diversity', 4, 0, 0, 3, 1],
['filmmaking', 0, 0, 0, 3, 0],
['said', 1272, 1049, 590, 700, 1631],
['bifa', 0, 0, 0, 1, 0],
['founder', 12, 14, 1, 5, 3],
['elliot', 0, 1, 0, 5, 1],
['grove', 0, 0, 0, 1, 0],
['commenting', 2, 0, 1, 1, 3],
['nominated', 3, 0, 0, 81, 2],
['added', 95, 80, 125, 79, 123],
['selection', 8, 0, 5, 3, 2],
['committee', 17, 8, 15, 12, 94],
['harder', 8, 0, 3, 2, 5],
['time', 249, 129, 231, 167, 204],
['ever', 42, 7, 28, 41, 49],
['narrowing', 0, 0, 0, 1, 0],

```
['field', 13, 12, 20, 14, 11],
['joining', 3, 3, 8, 6, 7],
['best', 41, 32, 101, 492, 31],
['climbing', 0, 1, 0, 1, 1],
['documentary', 2, 0, 0, 36, 7],
['touching', 2, 1, 1, 5, 0],
['void', 1, 0, 1, 3, 0],
['zombie', 4, 0, 0, 1, 0],
['comedy', 2, 0, 0, 112, 0],
['shaun', 1, 0, 6, 6, 1],
['geoffrey', 0, 0, 0, 3, 0],
['rush', 8, 0, 4, 6, 4],
['win', 22, 7, 270, 87, 58],
['actor', 5, 1, 0, 205, 1],
['role', 23, 21, 12, 121, 72],
['peter', 6, 15, 12, 15, 19],
['seller', 6, 1, 0, 9, 0],
['recent', 47, 79, 13, 33, 28],
['biopic', 0, 0, 0, 16, 0],
['life', 87, 33, 21, 100, 80],
['death', 9, 6, 3, 46, 26],
['australian', 4, 21, 56, 8, 5],
['star', 27, 5, 38, 260, 4],
['daniel', 0, 0, 4, 7, 1],
['craig', 1, 0, 13, 4, 0],
['phil', 3, 0, 16, 12, 3],
['davis', 0, 2, 27, 30, 28],
['ian', 11, 6, 8, 12, 9],
['hart', 0, 0, 11, 3, 2],
['blind', 16, 2, 0, 4, 0],
['fight', 17, 20, 17, 20, 39],
['paddy', 0, 0, 1, 3, 0],
['considine', 0, 0, 0, 3, 0],
['supporting', 2, 1, 0, 35, 4],
['rare', 4, 0, 4, 6, 2],
['u', 286, 501, 132, 300, 130],
['scarlett', 0, 0, 0, 3, 0],
['johansson', 0, 0, 13, 3, 0],
['among', 45, 39, 9, 63, 37],
['contender', 2, 0, 6, 9, 3],
['girl', 2, 0, 2, 22, 4],
['pearl', 1, 0, 0, 1, 0],
['earring', 0, 0, 0, 1, 0],
['fellow', 2, 2, 16, 22, 10],
['include', 46, 20, 5, 68, 25],
['imelda', 0, 0, 0, 24, 0],
['staunton', 0, 0, 0, 34, 0],
['natalie', 0, 0, 0, 11, 0],
['press', 19, 27, 19, 11, 25],
['anne', 0, 0, 0, 2, 3],
['reid', 0, 1, 2, 2, 3],
['mother', 5, 1, 1, 19, 7],
['eva', 0, 1, 0, 1, 0],
['birthistle', 0, 0, 0, 1, 0],
['ae', 0, 0, 0, 1, 0],
['fond', 2, 0, 1, 2, 1],
['kiss', 0, 0, 0, 1, 0],
['...', 15, 22, 10, 35, 36],
['kevin', 3, 1, 17, 14, 1],
['mcdonald', 0, 6, 0, 6, 1],
```

```
['former', 15, 69, 77, 56, 89],
['douglas', 0, 0, 11, 13, 4],
['hickox', 0, 0, 0, 1, 0],
['directorial', 0, 0, 0, 2, 0],
['debut', 13, 2, 17, 55, 2],
['seasoned', 0, 0, 0, 1, 0],
['film-makers', 0, 0, 0, 9, 0],
['roger', 2, 0, 14, 8, 8],
['michell', 0, 0, 0, 2, 0],
['pavel', 0, 0, 1, 1, 0],
['pavlikowsky', 0, 0, 0, 1, 0],
['challenge', 42, 9, 28, 12, 33],
['harry', 5, 0, 10, 34, 13],
['potter', 5, 0, 1, 27, 0],
['author', 10, 7, 0, 18, 3],
['jk', 0, 0, 0, 7, 0],
['rowling', 0, 0, 0, 7, 0],
['receive', 18, 13, 5, 17, 10],
['special', 19, 6, 8, 45, 24],
['contribution', 1, 10, 5, 12, 10],
['industry', 141, 96, 1, 76, 22],
['net', 262, 22, 16, 1, 6],
['regulation', 9, 7, 1, 4, 17],
['still', 100, 101, 95, 42, 91],
['possible', 42, 30, 25, 7, 44],
['blurring', 2, 0, 0, 0, 0],
['boundary', 3, 3, 0, 1, 1],
['internet', 197, 11, 0, 9, 10],
['raise', 16, 24, 5, 17, 28],
['question', 24, 11, 21, 10, 71],
['watchdog', 12, 12, 0, 7, 20],
['ofcom', 15, 8, 0, 3, 0],
['move', 78, 83, 50, 16, 76],
['closer', 10, 5, 4, 17, 7],
['year', 391, 553, 245, 491, 329],
['tv-quality', 2, 0, 0, 0, 0],
['video', 197, 9, 9, 35, 4],
['online', 201, 4, 0, 13, 7],
['becomes', 9, 6, 3, 5, 2],
['norm', 1, 0, 0, 0, 0],
['debate', 11, 4, 2, 9, 65],
['westminster', 1, 0, 0, 0, 31],
['industry', 141, 96, 1, 76, 22],
['considered', 21, 5, 6, 11, 6],
['option', 14, 22, 15, 0, 12],
['lord', 9, 7, 2, 16, 262],
['currie', 3, 0, 0, 0, 0],
['chairman', 13, 36, 27, 8, 60],
['super-regulator', 1, 0, 0, 0, 0],
['told', 102, 101, 102, 80, 257],
['panel', 21, 14, 5, 16, 10],
['protecting', 6, 2, 1, 1, 4],
['always', 45, 6, 39, 37, 38],
['primary', 4, 5, 2, 3, 3],
['concern', 30, 37, 8, 6, 72],
['despite', 45, 60, 47, 38, 35],
['remit', 1, 0, 0, 1, 1],
['disquiet', 1, 1, 0, 0, 2],
['increased', 11, 42, 3, 0, 22],
['among', 45, 39, 9, 63, 37],
```

```
['service', 391, 56, 10, 27, 170],
['provider', 31, 6, 0, 1, 0],
['speech', 38, 10, 1, 7, 52],
['made', 129, 97, 130, 119, 134],
['recent', 47, 79, 13, 33, 28],
['month', 162, 196, 81, 55, 103],
['hinted', 3, 2, 3, 1, 3],
['might', 57, 21, 26, 19, 39],
['organised', 13, 1, 2, 4, 8],
['association', 28, 19, 25, 12, 29],
['ispa', 1, 0, 0, 0, 0],
['possibility', 11, 7, 6, 4, 7],
['challenge', 42, 9, 28, 12, 33],
['arise', 2, 1, 0, 0, 3],
['truly', 9, 1, 1, 4, 3],
['blur', 7, 1, 1, 1, 0],
['balance', 4, 14, 0, 1, 11],
['struck', 6, 1, 5, 4, 5],
['consumer', 205, 71, 0, 2, 7],
['assess', 2, 7, 1, 0, 12],
['risk', 24, 55, 12, 5, 26],
['adopting', 4, 0, 0, 0, 1],
['currently', 86, 38, 12, 33, 32],
['exist', 10, 3, 1, 1, 0],
['regulate', 5, 0, 0, 2, 1],
['self-regulation', 1, 0, 0, 0, 0],
['practice', 13, 7, 1, 2, 13],
['discussion', 11, 8, 2, 4, 17],
['study', 44, 9, 2, 2, 23],
['suggest', 9, 11, 1, 2, 25],
['many', 282, 108, 46, 72, 116],
['eight', 21, 9, 35, 31, 20],
['million', 247, 97, 0, 125, 55],
['household', 17, 7, 0, 2, 14],
['uk', 197, 107, 5, 196, 232],
['could', 372, 196, 147, 75, 278],
['adopted', 11, 2, 0, 4, 5],
['broadband', 176, 6, 0, 0, 3],
['end', 94, 55, 73, 52, 70],
['2005', 109, 98, 18, 32, 22],
['open', 70, 22, 109, 27, 41],
['door', 12, 4, 12, 10, 19],
['delivered', 7, 8, 6, 2, 11],
['company', 272, 367, 4, 63, 22],
['streaming', 10, 0, 0, 0, 0],
['web', 141, 1, 0, 2, 0],
['already', 117, 47, 40, 51, 88],
['entertainment', 74, 3, 0, 30, 1],
['division', 10, 13, 3, 14, 8],
['distribute', 19, 1, 0, 2, 0],
['come', 111, 63, 99, 58, 114],
['source', 52, 15, 2, 4, 21],
['bskyb', 3, 4, 0, 1, 0],
['itv', 2, 0, 0, 31, 7],
['head', 41, 34, 28, 26, 24],
['andrew', 14, 3, 19, 12, 17],
['burke', 7, 0, 1, 0, 0],
['spoke', 1, 1, 5, 4, 9],
['creating', 19, 9, 4, 3, 12],
['platform', 14, 1, 3, 1, 12],
```

```
['risque', 1, 0, 0, 0, 0],
['celebrity', 9, 0, 0, 35, 4],
['chef', 1, 0, 0, 0, 1],
['serving', 1, 2, 2, 1, 6],
['expletive', 1, 0, 0, 1, 0],
['hot', 6, 0, 5, 17, 0],
['dinner', 2, 0, 2, 8, 6],
['surely', 3, 0, 0, 5, 6],
['push', 18, 10, 3, 2, 7],
['limit', 24, 13, 1, 6, 51],
['fact', 36, 22, 12, 21, 33],
['requested', 4, 0, 2, 1, 0],
['gone', 10, 10, 23, 17, 19],
['length', 2, 1, 0, 0, 4],
['download', 79, 0, 0, 33, 0],
['maybe', 7, 1, 16, 7, 6],
['entirely', 11, 4, 0, 1, 13],
['free', 94, 8, 11, 13, 31],
['long', 50, 38, 50, 33, 43],
['claimed', 16, 26, 26, 7, 40],
['responsibility', 12, 10, 7, 4, 24],
['carry', 27, 8, 7, 13, 8],
['server', 42, 0, 0, 0, 0],
['since', 68, 107, 96, 60, 67],
['law', 67, 51, 4, 17, 193],
['commission', 21, 40, 6, 3, 54],
['dubbed', 13, 2, 0, 4, 6],
['mere', 3, 1, 0, 0, 2],
['conduit', 1, 0, 0, 0, 0],
['back', 81, 88, 203, 55, 123],
['2002', 11, 31, 14, 38, 15],
['defence', 11, 14, 25, 4, 26],
['apply', 7, 2, 1, 0, 15],
['actual', 9, 1, 0, 1, 1],
['knowledge', 14, 1, 1, 3, 2],
['illegal', 32, 13, 3, 7, 25],
['failed', 14, 20, 31, 15, 45],
['remove', 17, 3, 0, 5, 8],
['level', 29, 74, 38, 8, 52],
['tested', 9, 0, 7, 4, 10],
['several', 76, 15, 24, 14, 21],
['high-profile', 10, 3, 2, 7, 4],
['legal', 85, 41, 3, 20, 72],
['case', 44, 76, 26, 22, 114],
['richard', 9, 7, 10, 31, 11],
['ayers', 2, 0, 0, 0, 0],
['portal', 18, 0, 0, 0, 0],
['director', 67, 51, 32, 190, 29],
['tiscali', 1, 0, 0, 0, 0],
['little', 48, 20, 36, 34, 46],
['point', 59, 53, 84, 22, 46],
['trying', 25, 28, 22, 20, 44],
['impossible', 7, 3, 3, 3, 8],
['huge', 60, 31, 22, 23, 25],
['change', 59, 50, 34, 26, 137],
['afoot', 1, 0, 0, 0, 1],
['predicted', 36, 17, 0, 8, 17],
['offer', 125, 103, 34, 8, 25],
['planned', 5, 19, 3, 12, 28],
['player', 243, 13, 261, 11, 4],
```

```
['give', 75, 41, 58, 28, 80],
['surfer', 12, 0, 0, 1, 0],
['chance', 24, 6, 113, 18, 42],
['programme', 72, 21, 7, 54, 78],
['eastenders', 3, 0, 0, 11, 0],
['top', 85, 34, 71, 146, 37],
['gear', 13, 1, 7, 2, 1],
['mainstream', 16, 0, 0, 14, 1],
['whole', 23, 18, 19, 14, 22],
['vast', 17, 4, 1, 1, 3],
['sum', 5, 1, 4, 3, 13],
['money', 63, 84, 45, 61, 78],
['maintaining', 6, 3, 1, 2, 5],
['network', 236, 21, 0, 26, 7],
['supply', 14, 33, 3, 2, 7],
['quantity', 4, 0, 0, 1, 0],
['data', 185, 45, 0, 2, 5],
['herald', 3, 0, 2, 0, 2],
['digital', 281, 1, 0, 36, 1],
['licence', 10, 2, 1, 7, 4],
['fee', 16, 8, 13, 5, 22],
['inappropriate', 4, 1, 0, 2, 3],
['obviously', 9, 3, 16, 8, 13],
['pornography', 7, 0, 0, 1, 0],
['viewed', 4, 4, 0, 1, 0],
['child', 35, 29, 1, 69, 93],
['continues', 11, 20, 5, 3, 10],
['dominate', 8, 4, 3, 0, 0],
['headline', 7, 7, 3, 7, 10],
['remains', 18, 36, 9, 12, 11],
['political', 22, 39, 1, 7, 107],
['issue', 56, 38, 22, 15, 193],
['mp', 1, 4, 0, 5, 231],
['allan', 2, 0, 3, 1, 1],
['liberal', 1, 1, 0, 2, 123],
['democrat', 2, 5, 0, 2, 112],
['spokesman', 36, 38, 16, 31, 138],
['think', 86, 30, 134, 93, 116],
['answer', 13, 1, 11, 5, 36],
['lie', 7, 1, 5, 6, 16],
['somewhere', 3, 2, 4, 2, 7],
['cry', 3, 0, 1, 10, 3],
['offline', 9, 0, 0, 0, 0],
['instead', 52, 18, 8, 12, 35],
['seeing', 22, 7, 9, 4, 6],
['brought', 20, 14, 18, 12, 29],
['future', 86, 71, 44, 15, 59],
['bring', 26, 12, 16, 12, 29],
['departed', 1, 0, 2, 0, 0],
['agreed', 18, 47, 13, 12, 31],
['reality', 13, 4, 4, 13, 8],
['power', 76, 21, 14, 14, 120],
['likely', 87, 70, 16, 11, 46],
['reign', 2, 0, 1, 1, 0],
['on-demand', 2, 0, 0, 0, 0],
['pulled', 2, 3, 14, 9, 1],
['rather', 62, 21, 10, 20, 50],
['pushed', 10, 17, 8, 3, 3],
['choice', 22, 5, 7, 29, 68],
['watershed', 1, 0, 0, 0, 2],
```

['format', 63, 2, 1, 8, 2],
['confuse', 4, 1, 0, 0, 2],
['firm', 325, 311, 9, 7, 25],
['sony', 129, 1, 0, 16, 0],
['philip', 8, 5, 1, 7, 2],
['matsushita', 4, 0, 0, 0, 0],
['samsung', 18, 0, 0, 0, 0],
['developing', 31, 18, 2, 6, 5],
['common', 29, 4, 6, 2, 112],
['way', 268, 57, 101, 56, 132],
['stop', 52, 13, 8, 15, 31],
['pirating', 2, 0, 0, 0, 0],
['want', 147, 44, 115, 49, 152],
['system', 279, 20, 3, 8, 102],
['ensures', 6, 1, 1, 0, 4],
['file', 125, 4, 2, 3, 1],
['play', 101, 9, 209, 101, 20],
['hardware', 23, 0, 0, 0, 0],
['thwart', 3, 0, 0, 0, 0],
['copying', 8, 1, 0, 1, 0],
['confusion', 4, 1, 1, 2, 2],
['faced', 12, 7, 2, 9, 18],
['different', 110, 8, 25, 29, 16],
['conflicting', 2, 0, 0, 0, 4],
['control', 104, 46, 19, 5, 60],
['expert', 49, 10, 1, 11, 23],
['warned', 29, 48, 11, 6, 53],
['say', 232, 122, 96, 86, 378],
['guarantee', 7, 3, 0, 1, 12],
['prevent', 14, 11, 3, 3, 13],
['piracy', 30, 2, 0, 7, 0],
['store', 53, 16, 0, 1, 3],
['wrap', 2, 0, 1, 0, 0],
['downloadable', 9, 0, 0, 0, 0],
['own-brand', 3, 0, 0, 0, 0],
['played', 17, 6, 95, 59, 7],
['small', 66, 19, 7, 15, 32],
['number', 208, 77, 82, 171, 98],
['known', 50, 18, 2, 28, 16],
['right', 86, 48, 98, 70, 221],
['management', 17, 32, 5, 5, 9],
['setting', 26, 3, 4, 3, 10],
['alliance', 12, 8, 0, 0, 21],
['hope', 36, 38, 52, 38, 47],
['current', 56, 52, 17, 19, 43],
['fragmentation', 2, 0, 0, 0, 0],
['joint', 12, 14, 3, 2, 10],
['statement', 36, 56, 30, 16, 49],
['wanted', 28, 13, 26, 30, 44],
['let', 109, 8, 17, 15, 30],
['enjoy', 13, 0, 5, 4, 2],
['appropriately', 2, 1, 0, 0, 1],
['licensed', 9, 1, 0, 0, 1],
['device', 219, 0, 0, 4, 6],
['independent', 16, 16, 9, 36, 30],
['obtained', 4, 0, 0, 0, 0],
['harder', 8, 0, 3, 2, 5],
['copy', 44, 2, 1, 31, 2],
['bought', 31, 36, 3, 8, 1],
['called', 78, 26, 23, 30, 76],

```
['marlin', 6, 0, 0, 0, 0],
['define', 3, 0, 1, 1, 1],
['basic', 18, 3, 0, 0, 16],
['specification', 11, 0, 0, 0, 0],
['every', 100, 16, 37, 27, 55],
['electronics', 63, 3, 0, 1, 0],
['conform', 4, 0, 0, 0, 0],
['built', 13, 17, 6, 5, 5],
['intertrust', 2, 0, 0, 0, 0],
['well', 118, 77, 121, 89, 82],
['earlier', 39, 60, 26, 26, 43],
['drm', 18, 0, 0, 0, 0],
['group', 99, 144, 32, 76, 93],
['coral', 2, 0, 0, 0, 0],
['consortium', 8, 11, 1, 0, 0],
['widely', 36, 19, 3, 9, 21],
['seen', 53, 57, 19, 42, 48],
['four', 51, 37, 91, 78, 57],
['decide', 7, 10, 15, 3, 18],
['destiny', 2, 2, 0, 5, 0],
['sign', 33, 23, 15, 8, 23],
['apple', 149, 0, 1, 3, 1],
['microsoft', 213, 5, 0, 0, 0],
['confusingly', 2, 0, 0, 0, 0],
['sit', 9, 0, 4, 1, 2],
['alongside', 16, 2, 13, 28, 4],
['rival', 37, 50, 26, 7, 8],
['akin', 5, 0, 0, 0, 0],
['physical', 11, 2, 3, 15, 1],
['betamax', 10, 0, 0, 0, 0],
['vhs', 6, 0, 0, 0, 0],
['past', 39, 47, 36, 35, 36],
['ian', 11, 6, 8, 12, 9],
['fogg', 15, 0, 0, 0, 0],
['analyst', 101, 162, 0, 0, 2],
['jupiter', 23, 0, 0, 0, 0],
['research', 142, 24, 0, 8, 29],
['difference', 22, 8, 10, 1, 22],
['fragmented', 2, 1, 1, 0, 0],
['two-horse', 2, 0, 2, 0, 0],
['race', 12, 2, 82, 13, 15],
['five', 58, 50, 88, 85, 42],
['six', 41, 28, 157, 37, 48],
['eight-horse', 2, 0, 0, 0, 0],
['careful', 7, 3, 1, 2, 3],
['buying', 23, 21, 0, 2, 7],
['ensure', 34, 15, 5, 12, 45],
['incompatibility', 3, 0, 0, 0, 0],
['within', 35, 24, 22, 16, 52],
['family', 25, 27, 9, 61, 67],
['although', 80, 47, 32, 22, 33],
['initiative', 21, 7, 3, 1, 21],
['sure', 31, 5, 34, 11, 36],
['program', 158, 3, 1, 0, 1],
['help', 143, 59, 26, 28, 80],
['uncertainty', 3, 11, 0, 1, 3],
['life', 87, 33, 21, 100, 80],
['confusing', 8, 0, 0, 0, 1],
['time', 249, 129, 231, 167, 204],
['shelley', 2, 0, 0, 0, 0],
```

```
['taylor', 5, 3, 23, 12, 11],
['author', 10, 7, 0, 18, 3],
['report', 118, 107, 30, 12, 141],
['lock', 3, 0, 23, 3, 1],
['done', 45, 8, 36, 25, 54],
['maximise', 2, 4, 0, 0, 1],
['cash', 66, 35, 9, 12, 23],
['itunes', 30, 0, 0, 7, 0],
['example', 34, 13, 8, 10, 30],
['hugely', 23, 1, 2, 4, 7],
['successful', 19, 11, 8, 19, 9],
['justify', 4, 1, 0, 0, 2],
['existence', 15, 0, 0, 1, 1],
['sell', 29, 26, 13, 10, 7],
['ipod', 81, 0, 0, 1, 0],
['rampant', 4, 1, 1, 0, 0],
['competition', 45, 26, 31, 25, 5],
['230', 3, 0, 0, 0, 1],
['figure', 60, 116, 9, 29, 58],
['drive', 77, 16, 16, 2, 11],
['openness', 2, 0, 0, 0, 1],
['freer', 2, 0, 0, 0, 0],
['win', 22, 7, 270, 87, 58],
['run', 58, 28, 56, 32, 36],
['listen', 26, 2, 7, 10, 7],
['earliest', 2, 0, 0, 2, 1],
['m', 57, 34, 0, 38, 68],
['place', 62, 23, 72, 81, 72],
['explain', 3, 2, 4, 3, 7],
['continuing', 15, 17, 4, 5, 10],
['popularity', 25, 3, 0, 6, 0],
['file-sharing', 39, 0, 0, 1, 0],
['get', 259, 45, 118, 88, 153],
['hold', 53, 35, 20, 8, 48],
['pirated', 25, 3, 0, 2, 0],
['pop', 5, 0, 0, 78, 2],
['portability', 7, 0, 0, 0, 0],
['peer-to-peer', 23, 0, 0, 0, 0],
['100', 43, 34, 6, 19, 32],
['cory', 3, 0, 0, 0, 0],
['doctorow', 6, 0, 0, 0, 0],
['electronic', 37, 5, 0, 6, 4],
['frontier', 6, 0, 0, 0, 0],
['foundation', 17, 5, 2, 7, 4],
['campaign', 55, 15, 14, 17, 157],
['cyber-rights', 2, 0, 0, 0, 0],
['expressed', 7, 7, 3, 5, 10],
['doubt', 19, 14, 32, 7, 23],
['achieve', 7, 2, 6, 0, 15],
['aim', 34, 10, 13, 4, 20],
['ever', 42, 7, 28, 41, 49],
['prevented', 5, 1, 3, 0, 2],
['readily', 2, 0, 0, 0, 1],
['admit', 3, 1, 2, 1, 1],
['protection', 25, 29, 0, 1, 13],
['skilled', 3, 0, 0, 0, 0],
['attacker', 4, 0, 1, 0, 4],
['crime', 22, 3, 0, 12, 50],
['gang', 9, 1, 0, 5, 0],
['responsible', 11, 10, 0, 15, 19],
```

```
['intended', 10, 3, 4, 5, 12],
['studio', 51, 1, 0, 54, 3],
['label', 8, 9, 0, 37, 3],
['perceive', 5, 0, 0, 0, 0],
['opportunity', 29, 17, 34, 12, 37],
['auto', 10, 11, 0, 0, 0],
['american', 40, 27, 41, 58, 11],
['ringtone', 8, 0, 0, 0, 0],
['russian', 7, 66, 20, 8, 1],
['bbc', 115, 37, 68, 149, 190],
['world', 212, 175, 223, 105, 96],
['prize', 17, 2, 16, 103, 4],
['return', 22, 26, 70, 41, 38],
['vozvrashchenie', 0, 0, 0, 2, 0],
['named', 4, 3, 16, 58, 3],
['four', 51, 37, 91, 78, 57],
['cinema', 10, 0, 0, 45, 1],
['tell', 26, 4, 9, 17, 17],
['story', 33, 4, 10, 56, 14],
['two', 154, 146, 214, 171, 154],
['adolescent', 1, 0, 0, 2, 0],
['boy', 3, 0, 10, 62, 1],
['subjected', 0, 0, 1, 2, 0],
['harsh', 2, 2, 2, 2, 0],
['regime', 5, 8, 2, 5, 10],
['strict', 4, 3, 0, 3, 2],
['father', 1, 3, 0, 21, 21],
['10', 91, 70, 63, 77, 51],
['absence', 0, 1, 9, 5, 4],
['directed', 0, 3, 0, 41, 1],
['andrey', 0, 0, 0, 2, 0],
['zvyagintsev', 0, 0, 0, 2, 0],
['previously', 5, 30, 5, 27, 8],
['2003', 44, 96, 35, 54, 35],
['golden', 5, 12, 3, 47, 8],
['lion', 3, 1, 50, 5, 0],
['festival', 0, 0, 0, 114, 0],
['presented', 9, 7, 6, 19, 6],
['held', 25, 27, 27, 46, 65],
['thursday', 5, 43, 12, 24, 46],
['hosted', 2, 1, 2, 12, 7],
['jonathan', 5, 1, 14, 11, 8],
['ross', 7, 1, 5, 24, 0],
['panel', 21, 14, 5, 16, 10],
['included', 19, 11, 10, 51, 10],
['x', 0, 1, 0, 10, 0],
['file', 125, 4, 2, 3, 1],
['gillian', 0, 0, 1, 2, 0],
['anderson', 0, 0, 4, 3, 0],
['critic', 8, 12, 5, 45, 29],
['clarke', 0, 3, 2, 3, 68],
['presenter', 0, 0, 0, 31, 2],
['one', 388, 149, 206, 289, 229],
['2005', 109, 98, 18, 32, 22],
['involved', 31, 18, 19, 25, 28],
['deliberation', 0, 0, 1, 2, 2],
['shortlist', 0, 0, 0, 16, 12],
['six', 41, 28, 157, 37, 48],
['around', 132, 38, 18, 54, 43],
['drawn', 3, 6, 9, 9, 10],
```

['chose', 1, 0, 5, 7, 4],
['motorcycle', 3, 0, 10, 10, 0],
['diary', 7, 0, 0, 20, 4],
['zatoichi', 0, 0, 0, 2, 0],
['hero', 6, 0, 9, 21, 0],
['viewer', 34, 0, 0, 44, 0],
['poll', 1, 11, 7, 17, 64],
['saw', 14, 22, 31, 44, 15],
['zhang', 1, 0, 0, 7, 0],
['yimou', 0, 0, 0, 3, 0],
['martial', 0, 0, 0, 2, 0],
['art', 24, 2, 0, 34, 13],
['epic', 1, 0, 3, 14, 0],
['emerge', 0, 7, 2, 4, 1],
['favourite', 18, 1, 20, 45, 2],
['32', 4, 6, 10, 2, 4],
['vote', 4, 5, 3, 33, 130],
['cast', 8, 3, 5, 26, 2],
['tragedy', 0, 6, 4, 6, 3],
['struck', 6, 1, 5, 4, 5],
['production', 22, 71, 0, 56, 0],
['young', 10, 8, 27, 63, 49],
['15', 32, 36, 11, 15, 16],
['year-old', 7, 6, 108, 52, 19],
['vladimir', 0, 7, 0, 2, 1],
['girin', 0, 0, 0, 2, 0],
['drowned', 1, 0, 0, 3, 0],
['lake', 0, 2, 0, 2, 1],
['last', 138, 226, 219, 211, 153],
['french', 10, 35, 46, 36, 10],
['animated', 1, 0, 0, 16, 0],
['feature', 50, 3, 5, 40, 6],
['belleville', 0, 0, 0, 2, 0],
['rendezvous', 0, 0, 0, 2, 0],
['rapper', 0, 0, 0, 30, 0],
['50', 42, 22, 6, 55, 18],
['cent', 3, 11, 0, 31, 3],
['end', 94, 55, 73, 52, 70],
['protege', 0, 0, 1, 6, 0],
['feud', 0, 3, 2, 8, 8],
['ended', 3, 11, 20, 12, 7],
['public', 46, 60, 13, 58, 224],
['game', 520, 11, 380, 40, 9],
['pair', 4, 5, 27, 21, 18],
['wanted', 28, 13, 26, 30, 44],
['good', 108, 69, 148, 70, 81],
['model', 70, 22, 2, 8, 1],
['community', 36, 9, 1, 17, 48],
['row', 1, 14, 28, 11, 21],
['blew', 2, 0, 3, 2, 1],
['threw', 1, 2, 8, 3, 2],
['g-unit', 0, 0, 0, 5, 0],
['crew', 2, 2, 0, 7, 0],
['accused', 7, 26, 8, 10, 53],
['disloyal', 0, 0, 0, 2, 0],
['member', 27, 46, 16, 49, 113],
['entourage', 0, 1, 0, 4, 0],
['reportedly', 2, 10, 4, 6, 3],
['shot', 9, 2, 43, 29, 6],
['outside', 17, 8, 16, 32, 21],

```
['radio', 118, 5, 18, 74, 63],
['station', 41, 4, 2, 36, 3],
['interviewed', 2, 2, 0, 2, 6],
['shook', 1, 0, 2, 3, 0],
['hand', 33, 26, 24, 11, 28],
['handed', 12, 2, 14, 14, 5],
['money', 63, 84, 45, 61, 78],
['music', 322, 3, 0, 378, 1],
['project', 91, 28, 0, 33, 20],
['new', 402, 253, 151, 238, 328],
['york', 22, 33, 12, 47, 4],
['deprived', 2, 0, 0, 2, 3],
['area', 52, 45, 13, 12, 53],
['wednesday', 7, 40, 44, 24, 46],
['whose', 5, 20, 12, 31, 18],
['real', 60, 18, 56, 36, 44],
['name', 70, 16, 18, 52, 22],
['jayceon', 0, 0, 0, 1, 0],
['taylor', 5, 3, 23, 12, 11],
['told', 102, 101, 102, 80, 257],
['news', 139, 98, 20, 42, 82],
['conference', 34, 10, 11, 6, 71],
['want', 147, 44, 115, 49, 152],
['apologise', 0, 1, 13, 3, 13],
['almost', 63, 38, 19, 21, 34],
['ashamed', 0, 0, 1, 2, 5],
['participated', 0, 3, 0, 1, 0],
['thing', 78, 19, 66, 45, 70],
['went', 28, 15, 42, 59, 43],
['week', 81, 96, 127, 102, 144],
['chart-topper', 0, 0, 0, 5, 0],
['curtis', 0, 0, 1, 11, 0],
['jackson', 1, 0, 3, 45, 6],
['truce', 0, 2, 0, 2, 1],
['came', 31, 50, 71, 57, 42],
['anniversary', 5, 0, 0, 15, 6],
['notorious', 1, 0, 1, 1, 1],
['big', 72, 30, 54, 68, 48],
['1997', 3, 5, 6, 9, 29],
['part', 116, 92, 48, 62, 114],
['volatile', 0, 2, 0, 1, 0],
['east', 0, 18, 6, 7, 46],
['west', 3, 13, 25, 48, 27],
['coast', 1, 2, 2, 2, 8],
['rap', 9, 0, 1, 22, 0],
['scene', 14, 0, 3, 38, 7],
['today', 28, 15, 13, 19, 62],
['show', 155, 32, 12, 323, 59],
['people', 726, 118, 58, 181, 451],
['rise', 25, 128, 4, 14, 56],
['difficult', 29, 19, 34, 10, 26],
['circumstance', 1, 4, 3, 1, 11],
['together', 52, 12, 16, 17, 22],
['put', 76, 61, 85, 32, 86],
['negativity', 0, 0, 0, 1, 0],
['behind', 60, 16, 31, 32, 22],
['lot', 74, 18, 94, 32, 24],
['see', 133, 56, 76, 59, 92],
['happen', 13, 4, 24, 8, 19],
['responding', 0, 2, 2, 1, 11],
```

```
['important', 57, 23, 34, 22, 60],
['group', 99, 144, 32, 76, 93],
['family', 25, 27, 9, 61, 67],
['fan', 54, 5, 44, 78, 1],
['choir', 0, 0, 0, 5, 1],
['harlem', 0, 0, 0, 5, 0],
['got', 35, 7, 108, 49, 53],
['cheque', 2, 0, 2, 2, 0],
['000', 142, 161, 24, 106, 150],
['£', 125, 289, 69, 198, 286],
['77', 0, 1, 7, 2, 0],
['800', 6, 7, 0, 3, 10],
['500', 20, 16, 2, 7, 15],
['53', 2, 4, 4, 3, 3],
['400', 3, 13, 0, 5, 7],
['made', 129, 97, 130, 119, 134],
['compton', 0, 0, 0, 1, 0],
['school', 30, 6, 5, 55, 106],
['programme', 72, 21, 7, 54, 78],
['launched', 62, 18, 4, 10, 21],
['g-unity', 0, 0, 0, 1, 0],
['foundation', 17, 5, 2, 7, 4],
['help', 143, 59, 26, 28, 80],
['overcome', 14, 2, 3, 3, 2],
['obstacle', 1, 1, 1, 1, 0],
['make', 294, 78, 119, 93, 169],
['chance', 24, 6, 113, 18, 42],
['better', 61, 30, 53, 24, 47],
['realised', 5, 1, 2, 8, 3],
['going', 110, 36, 110, 66, 114],
['effective', 7, 4, 4, 2, 17],
['need', 120, 72, 53, 26, 122],
['set', 125, 72, 137, 71, 99],
['example', 34, 13, 8, 10, 30],
['stranger', 1, 0, 1, 2, 0],
['ja', 0, 0, 0, 4, 0],
['rule', 37, 45, 13, 11, 91],
['target', 23, 49, 11, 4, 59],
['ridicule', 0, 0, 0, 1, 0],
['song', 41, 1, 0, 206, 2],
...]
```

Observation:

- converted all the word frequencies in [word, class1_frequency, class2_frequency, class3_frequency, class4_frequency, class5_frequency] form.

In [269…

```python
def extract_features(word_list, freqs):
    '''
        tweet: a list of words for one tweet
        freqs: a dictionary corresponding to the frequencies of each tuple (word
        x: a feature vector of dimension (1,3)
    '''

    # 5 elements in the form of a 1 x 6 vector
    x = np.zeros((1, 6))

    #bias term is set to 1
    x[0,0] = 1
```

```
    # loop through each word in the list of words
    for word in word_list:

        # increment the word count for the Technology label 1
        x[0,1] += freqs.get((word, 1),0)

        # increment the word count for the Business label 2
        x[0,2] += freqs.get((word, 2),0)

        # increment the word count for the Business label 3
        x[0,3] += freqs.get((word, 3),0)

        # increment the word count for the Business label 4
        x[0,4] += freqs.get((word, 4),0)

        # increment the word count for the Business label 5
        x[0,5] += freqs.get((word, 5),0)

    assert(x.shape == (1, 6))
    return x
```

In [270…

```
df_train_shuffled
```

Out[270…

	Article_Cleaned	Category
0	[viewer, able, shape, tv, imagine, editing, ti...	1
1	[indie, film, nomination, announced, mike, lei...	4
2	[net, regulation, still, possible, blurring, b...	1
3	[format, war, could, confuse, user, technology...	1
4	[russian, film, win, bbc, world, prize, russia...	4
...	...	...
1535	[potter, director, sign, warner, deal, harry, ...	4
1536	[fox, attack, blair, tory, lie, tony, blair, l...	5
1537	[capriati, miss, melbourne, jennifer, capriati...	3
1538	[brown, name, 16, march, budget, chancellor, g...	5
1539	[cage, film, third, week, u, top, nicolas, cag...	4

1540 rows × 2 columns

Observation: Shuffled the training data succesfully.

Preparing training data

In [271…

```
# collect the features 'x' and stack them into a matrix 'X'
X_train_lr = np.zeros((len(df_train_shuffled), 6))
for i in range(len(df_train_shuffled)):
    X_train_lr[i, :]= extract_features(df_train_shuffled['Article_Cleaned'][i],
```

```
# training labels corresponding to X
y_train_lr = np.array(df_train_shuffled['Category'].to_list())
# Y = np.ravel(train_y,order='C')
```

In [272…
```
X_test_lr = np.zeros((len(df_test_shuffled), 6))
for i in range(len(df_test_shuffled)):
    X_test_lr[i, :]= extract_features(df_test_shuffled['Article_Cleaned'][i], fi

y_test_lr = np.array(df_test_shuffled['Category'].to_list())
```

In [273…
```
X_train_lr
```

Out[273…
```
array([[1.0000e+00, 2.7042e+04, 1.2986e+04, 9.1170e+03, 1.6088e+04,
        1.9938e+04],
       [1.0000e+00, 7.6230e+03, 6.4220e+03, 5.6970e+03, 2.0891e+04,
        6.8680e+03],
       [1.0000e+00, 4.1370e+04, 2.0928e+04, 1.2457e+04, 1.6060e+04,
        3.4147e+04],
       ...,
       [1.0000e+00, 3.9310e+03, 3.4440e+03, 4.5350e+03, 3.3580e+03,
        4.0540e+03],
       [1.0000e+00, 7.8500e+03, 9.3460e+03, 4.0300e+03, 4.4060e+03,
        2.0982e+04],
       [1.0000e+00, 4.6160e+03, 4.6090e+03, 3.3600e+03, 7.4880e+03,
        4.3720e+03]])
```

In [274…
```
y_train_lr
```

Out[274…
```
array([1, 4, 1, ..., 3, 5, 4])
```

In [275…
```
X_test_lr
```

Out[275…
```
array([[1.0000e+00, 1.0819e+04, 1.0495e+04, 5.0230e+03, 6.2630e+03,
        1.0938e+04],
       [1.0000e+00, 6.5580e+03, 5.3460e+03, 3.5450e+03, 5.9270e+03,
        6.5090e+03],
       [1.0000e+00, 2.5783e+04, 2.0452e+04, 1.3462e+04, 1.6446e+04,
        2.6941e+04],
       ...,
       [1.0000e+00, 1.8743e+04, 1.9130e+04, 7.9070e+03, 9.8290e+03,
        1.6078e+04],
       [1.0000e+00, 1.4692e+04, 9.8440e+03, 1.3396e+04, 9.5860e+03,
        1.4756e+04],
       [1.0000e+00, 1.2138e+04, 8.7080e+03, 3.6750e+03, 5.4330e+03,
        8.7600e+03]])
```

In [276…
```
y_test_lr
```

Out[276…

```
array([2, 4, 2, 3, 2, 3, 2, 3, 2, 5, 4, 3, 2, 1, 2, 2, 2, 1, 3, 1, 2, 5,
       2, 2, 2, 2, 4, 3, 2, 3, 3, 3, 3, 1, 1, 5, 4, 5, 4, 1, 2, 4, 5, 1,
       2, 2, 3, 2, 3, 1, 3, 2, 3, 3, 2, 2, 1, 3, 3, 5, 2, 2, 1, 2, 5, 1,
       3, 4, 3, 5, 3, 2, 1, 2, 2, 3, 1, 5, 1, 1, 3, 5, 1, 1, 2, 4, 2, 2,
       5, 4, 3, 4, 1, 3, 1, 2, 2, 1, 3, 3, 1, 2, 4, 4, 4, 5, 2, 3, 3, 5,
       3, 1, 3, 2, 3, 5, 2, 3, 1, 4, 2, 2, 1, 5, 3, 5, 3, 3, 4, 1, 5, 2,
       2, 2, 1, 1, 2, 4, 4, 3, 2, 2, 1, 5, 3, 5, 2, 1, 2, 3, 1, 2, 5, 4,
       3, 3, 4, 3, 3, 5, 2, 3, 3, 2, 2, 2, 4, 5, 2, 5, 5, 3, 3, 3, 5, 2,
       2, 5, 2, 2, 1, 1, 1, 5, 4, 4, 3, 3, 3, 2, 3, 1, 1, 1, 2, 5, 5, 1,
       1, 2, 4, 4, 5, 3, 4, 5, 3, 2, 3, 3, 1, 4, 3, 3, 1, 1, 1, 3, 4, 3,
       2, 5, 3, 2, 3, 3, 1, 2, 2, 1, 1, 3, 2, 5, 5, 3, 1, 3, 3, 5, 2, 2,
       2, 3, 5, 2, 5, 5, 2, 3, 5, 1, 2, 4, 5, 2, 3, 2, 1, 4, 1, 3, 2, 2,
       1, 5, 2, 2, 5, 2, 1, 5, 2, 3, 3, 2, 2, 3, 3, 2, 3, 2, 5, 4, 5, 3,
       2, 3, 1, 2, 2, 5, 4, 2, 3, 4, 3, 2, 3, 5, 5, 2, 3, 2, 2, 4, 5, 2,
       3, 3, 5, 3, 5, 2, 2, 3, 1, 2, 5, 3, 5, 1, 3, 1, 4, 2, 3, 3, 2, 3,
       3, 2, 2, 2, 5, 2, 2, 4, 2, 3, 1, 3, 2, 2, 1, 3, 3, 2, 3, 5, 2, 5,
       5, 3, 2, 2, 3, 4, 3, 3, 3, 3, 1, 1, 3, 4, 4, 1, 1, 5, 5, 1, 3, 4,
       3, 5, 2, 3, 2, 3, 2, 1, 4, 2, 1, 3, 2, 1, 2, 2, 5, 3, 3, 2, 5, 5,
       2, 5, 5, 5, 3, 5, 4, 5, 1, 3, 2, 5, 2, 3, 2, 3, 3, 3, 5, 3, 2, 5,
       5, 2, 3, 4, 2, 3, 3, 4, 1, 2, 3, 5, 2, 1, 2, 3, 5, 2, 3, 3, 4, 4,
       1, 4, 1, 2, 2, 3, 4, 4, 3, 1, 2, 5, 5, 3, 2, 4, 3, 3, 2, 3, 3, 2,
       5, 3, 5, 1, 2, 3, 3, 3, 2, 3, 3, 2, 3, 5, 4, 3, 3, 5, 2, 4, 3, 3,
       3, 2, 2, 5, 2, 2, 2, 3, 3, 5, 2, 3, 3, 2, 2, 2, 3, 2, 2, 2, 3, 3,
       2, 3, 2, 5, 4, 2, 5, 2, 3, 4, 4, 2, 5, 4, 5, 4, 3, 5, 4, 2, 3, 3,
       2, 2, 5, 3, 2, 4, 3, 3, 5, 3, 2, 2, 3, 3, 2, 5, 4, 3, 3, 4, 5, 5,
       5, 1, 2, 2, 4, 1, 4, 2, 5, 5, 5, 2, 2, 3, 4, 5, 2, 3, 3, 5, 2, 1,
       5, 4, 2, 4, 3, 4, 1, 1, 2, 2, 1, 5, 2, 3, 3, 2, 3, 3, 4, 2, 1, 3,
       2, 5, 4, 1, 1, 2, 1, 3, 3, 2, 2, 1, 3, 3, 2, 3, 3, 5, 2, 2, 2, 4,
       3, 3, 2, 4, 3, 1, 5, 4, 1, 2, 1, 4, 3, 3, 3, 1, 3, 4, 3, 2, 1, 3,
       5, 3, 2, 2, 5, 3, 5, 4, 2, 4, 4, 2, 3, 4, 1, 3, 4, 2, 1, 3, 3, 2,
       2, 2, 3, 3, 3, 3, 3, 5, 2, 5, 2, 3, 1, 3, 3, 2, 3, 2, 5, 1, 1, 2,
       2, 3, 2])
```

Obseravtion:

Extracted the X_train_lr, y_train_lr, X_test_lr, y_test_lr data successfully.

Preprocessing Functions Declaration for the following models

- Naive Bayes
- KNN Classifier
- Random Forest
- Decision Tree Classifer

In [277…

```
df
```

Out[277…

	Category	Article	Encoded_Cat	Article_Cleaned
0	Technology	tv future in the hands of viewers with home th...	1	[tv, future, hand, viewer, home, theatre, syst...
1	Business	worldcom boss left books alone former worldc...	2	[worldcom, bos, left, book, alone, former, wor...
2	Sports	tigers wary of farrell gamble leicester say ...	3	[tiger, wary, farrell, gamble, leicester, say,...
3	Sports	yeading face newcastle in fa cup premiership s...	3	[yeading, face, newcastle, fa, cup, premiershi...
4	Entertainment	ocean s twelve raids box office ocean s twelve...	4	[ocean, twelve, raid, box, office, ocean, twel...
...	...	...	...	...
2220	Business	cars pull down us retail figures us retail sal...	2	[car, pull, u, retail, figure, u, retail, sale...
2221	Politics	kilroy unveils immigration policy ex-chatshow ...	5	[kilroy, unveils, immigration, policy, ex-chat...
2222	Entertainment	rem announce new glasgow concert us band rem h...	4	[rem, announce, new, glasgow, concert, u, band...
2223	Politics	how political squabbles snowball it s become c...	5	[political, squabble, snowball, become, common...
2224	Sports	souness delight at euro progress boss graeme s...	3	[souness, delight, euro, progress, bos, graeme...

2225 rows × 4 columns

In [278…

```
df['Article'][0]
```

Out[278…

'tv future in the hands of viewers with home theatre systems plasma high-defin
ition tvs and digital video recorders moving into the living room the way peo
ple watch tv will be radically different in five years time. that is accordin
g to an expert panel which gathered at the annual consumer electronics show in
las vegas to discuss how these new technologies will impact one of our favourit
e pastimes. with the us leading the trend programmes and other content will be
delivered to viewers via home networks through cable satellite telecoms comp
anies and broadband service providers to front rooms and portable devices. on
e of the most talked-about technologies of ces has been digital and personal vi
deo recorders (dvr and pvr). these set-top boxes like the us s tivo and the uk
s sky+ system allow people to record store play pause and forward wind tv p
rogrammes when they want. essentially the technology allows for much more per
sonalised tv. they are also being built-in to high-definition tv sets which ar
e big business in japan and the us but slower to take off in europe because of
the lack of high-definition programming. not only can people forward wind throu
gh adverts they can also forget about abiding by network and channel schedules
putting together their own a-la-carte entertainment. but some us networks and c
able and satellite companies are worried about what it means for them in terms
of advertising revenues as well as brand identity and viewer loyalty to chann
els. although the us leads in this technology at the moment it is also a conce
rn that is being raised in europe particularly with the growing uptake of serv
ices like sky+. what happens here today we will see in nine months to a years
time in the uk adam hume the bbc broadcast s futurologist told the bbc news
website. for the likes of the bbc there are no issues of lost advertising reve
nue yet. it is a more pressing issue at the moment for commercial uk broadcaste
rs but brand loyalty is important for everyone. we will be talking more about
content brands rather than network brands said tim hanlon from brand communi
cations firm starcom mediavest. the reality is that with broadband connections
anybody can be the producer of content. he added: the challenge now is that i
t is hard to promote a programme with so much choice. what this means said s
tacey jolna senior vice president of tv guide tv group is that the way people
find the content they want to watch has to be simplified for tv viewers. it mea
ns that networks in us terms or channels could take a leaf out of google s bo
ok and be the search engine of the future instead of the scheduler to help peo
ple find what they want to watch. this kind of channel model might work for the
younger ipod generation which is used to taking control of their gadgets and wh
at they play on them. but it might not suit everyone the panel recognised. old
er generations are more comfortable with familiar schedules and channel brands
because they know what they are getting. they perhaps do not want so much of th
e choice put into their hands mr hanlon suggested. on the other end you have
the kids just out of diapers who are pushing buttons already - everything is po
ssible and available to them said mr hanlon. ultimately the consumer will t
ell the market they want. of the 50 000 new gadgets and technologies being sh
owcased at ces many of them are about enhancing the tv-watching experience. hi
gh-definition tv sets are everywhere and many new models of lcd (liquid crystal
display) tvs have been launched with dvr capability built into them instead of
being external boxes. one such example launched at the show is humax s 26-inch
lcd tv with an 80-hour tivo dvr and dvd recorder. one of the us s biggest satel
lite tv companies directtv has even launched its own branded dvr at the show
with 100-hours of recording capability instant replay and a search function.
the set can pause and rewind tv for up to 90 hours. and microsoft chief bill ga
tes announced in his pre-show keynote speech a partnership with tivo called ti
votogo which means people can play recorded programmes on windows pcs and mobi
le devices. all these reflect the increasing trend of freeing up multimedia so
that people can watch what they want when they want.'

In [279…

```
def covert_to_string(word_list):
  strn = ''
  for word in word_list:
    strn = strn + ' ' + word
```

```
  strn = strn.strip()
  return strn
```

In [280…

```
def cv_tfidf(df,X,y, choice):

  if choice == 1:
    cv = CountVectorizer(max_features=5000)
    X = cv.fit_transform(df[X].apply(covert_to_string)).toarray()
    y = np.array(df[y].values)

  elif choice == 2:
    tf_idf = TfidfVectorizer()
    X = tf_idf.fit_transform(df[X].apply(covert_to_string)).toarray()
    y = np.array(df[y].values)

  else:
    print("Wrong Input!")

  return X,y
```

In [281…

```
def train_n_get_metrics(model, X_train, y_train, X_test, y_test):
  # Calculating the train & test accuracy -
  model.fit(X_train, y_train)

  model_details = {}

  print(f"{'*' * 100}")
  model_name = type(model).__name__
  print("Model Name : {}\n".format(model_name))
  y_test_pred_nb = model.predict(X_test)
  y_train_pred_nb = model.predict(X_train)

  nb_train = accuracy_score(y_train, y_train_pred_nb)
  nb_test = accuracy_score(y_test, y_test_pred_nb)

  print("Train accuracy :{:.3f}".format(nb_train))
  print("Test accuracy :{:.3f}\n".format(nb_test))

  # Making predictions on the test set -
  y_pred_proba_nb = model.predict_proba(X_test)
  _roc_auc_score = roc_auc_score(y_test, y_pred_proba_nb, multi_class='ovr')
  print("ROC AUC Score: {:.3f}\n".format(_roc_auc_score))

  precision = precision_score(y_test, y_test_pred_nb, average='weighted')
  recall = recall_score(y_test, y_test_pred_nb, average='weighted')
  f1 = f1_score(y_test, y_test_pred_nb, average='weighted')

  print("Precision: {:.3f}".format(precision))
  print("Recall: {:.3f}".format(recall))
  print("F1 Score: {:.3f}\n".format(f1))

  print(classification_report(y_test, y_test_pred_nb))

  cm = confusion_matrix(y_test, y_test_pred_nb)

  plt.figure(figsize = (8, 5))
  sns.heatmap(cm, annot=True, fmt='d', cbar=False, cmap='Blues')
  plt.title('Confusion Matrix')
  plt.xlabel('Predicted label')
```

```
  plt.ylabel('Actual label')
  plt.show()
  print(f"{'*' * 100}")

  model_details[model_name] = {"Train_accuracy" : round(nb_train,3),
                               "Test_accuracy" : round(nb_test,3),
                               "ROC AUC Score" : round(_roc_auc_score,3),
                               "Precision" : round(precision,3),
                               "Recall" : round(recall,3),
                               "F1_score" : round(f1,3)}

  return model_details
```

Modelling Manual Vectorization

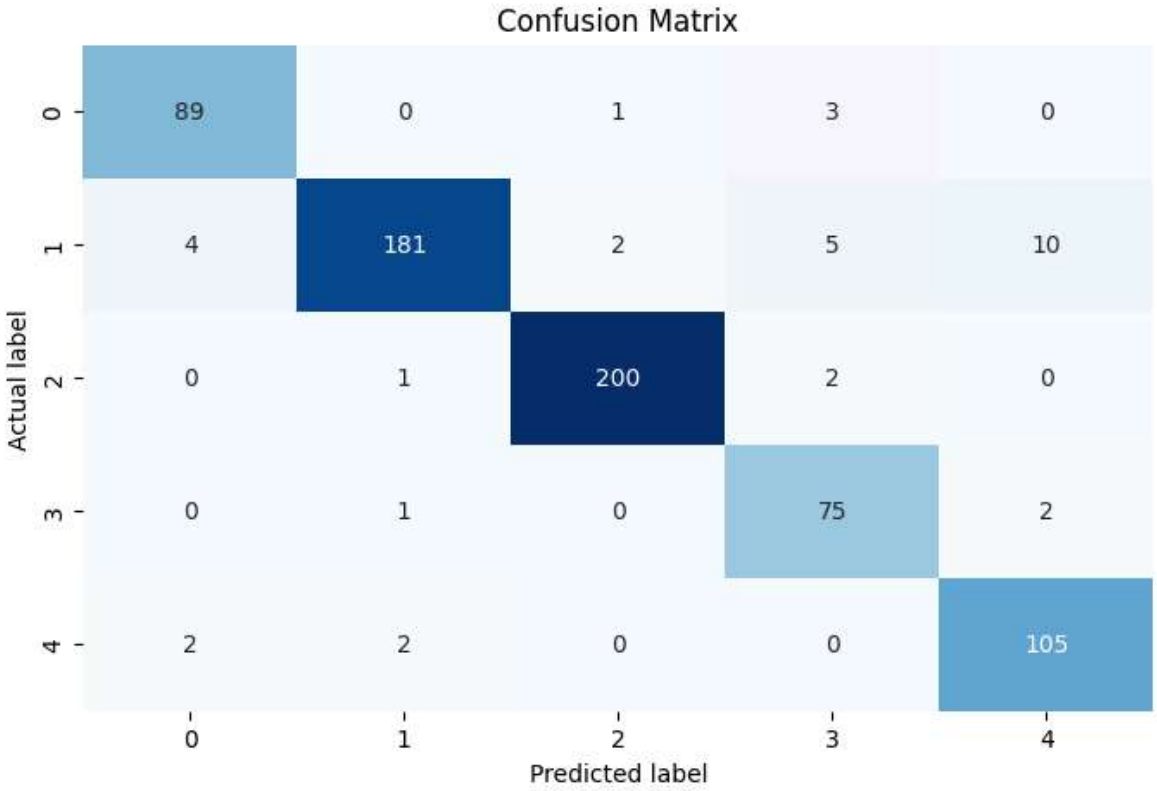
Logistic Regression (Manually engineered statistical features with bias term)

In [282…

```
lr = LogisticRegression(multi_class='ovr', solver='liblinear')
lr_details = train_n_get_metrics(lr, X_train_lr, y_train_lr, X_test_lr, y_test_l
```

```
**********************************************************************************
******************
Model Name : LogisticRegression

Train accuracy :0.964
Test accuracy :0.949

ROC AUC Score: 0.995

Precision: 0.951
Recall: 0.949
F1 Score: 0.949

              precision    recall  f1-score   support

           1       0.94      0.96      0.95        93
           2       0.98      0.90      0.94       202
           3       0.99      0.99      0.99       203
           4       0.88      0.96      0.92        78
           5       0.90      0.96      0.93       109

    accuracy                           0.95       685
   macro avg       0.94      0.95      0.94       685
weighted avg       0.95      0.95      0.95       685
```

```
/usr/local/lib/python3.11/dist-packages/sklearn/linear_model/_logistic.py:1256: F
utureWarning:

'multi_class' was deprecated in version 1.5 and will be removed in 1.7. Use OneVs
RestClassifier(LogisticRegression(..)) instead. Leave it to its default value to
avoid this warning.
```



****

Modeling using BOW

In [283…
```
X,y = cv_tfidf(df, 'Article_Cleaned', 'Encoded_Cat', 1)
```

In [284…
```
X
```

Out[284…
```
array([[1, 0, 0, ..., 0, 0, 0],
       [1, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [1, 0, 0, ..., 0, 0, 0],
       [1, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

In [285…
```
y
```

Out[285…
```
array([1, 2, 3, ..., 4, 5, 3])
```

In [286…
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,
                                                    shuffle=True, stratify=y,
                                                    random_state=42)
```

Observation:

- Processed the X_train, X_test, y_train, y_test using BOW technique.

In [287…
```
print("No. of rows in train set is {}.".format(X_train.shape[0]))
print("No. of rows in test set is {}.".format(X_test.shape[0]))
```

```
No. of rows in train set is 1668.
No. of rows in test set is 557.
```

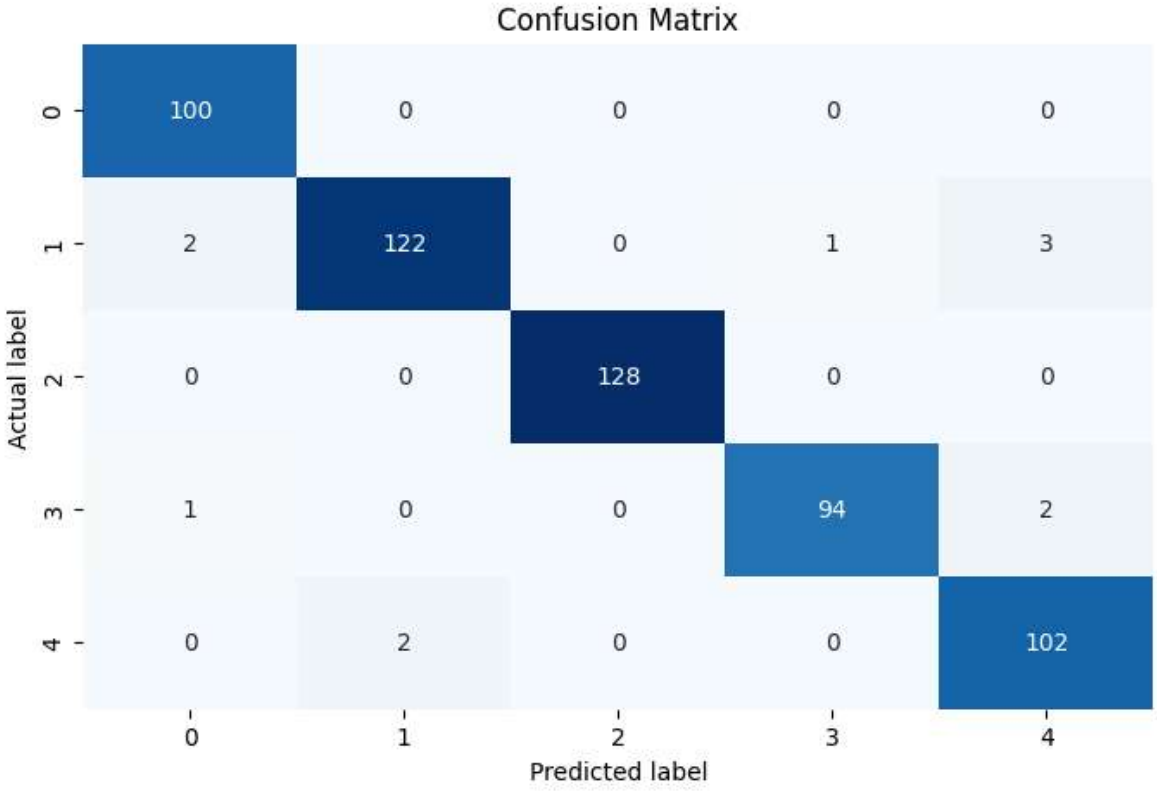
Naive Bayes Classifier

In [288…

```
bow_mnb_details = train_n_get_metrics(MultinomialNB(), X_train, y_train, X_test,
```

```
**********************************************************************************
******************
Model Name : MultinomialNB

Train accuracy :0.989
Test accuracy :0.980

ROC AUC Score: 0.998

Precision: 0.981
Recall: 0.980
F1 Score: 0.980
```

	precision	recall	f1-score	support
1	0.97	1.00	0.99	100
2	0.98	0.95	0.97	128
3	1.00	1.00	1.00	128
4	0.99	0.97	0.98	97
5	0.95	0.98	0.97	104
accuracy			0.98	557
macro avg	0.98	0.98	0.98	557
weighted avg	0.98	0.98	0.98	557



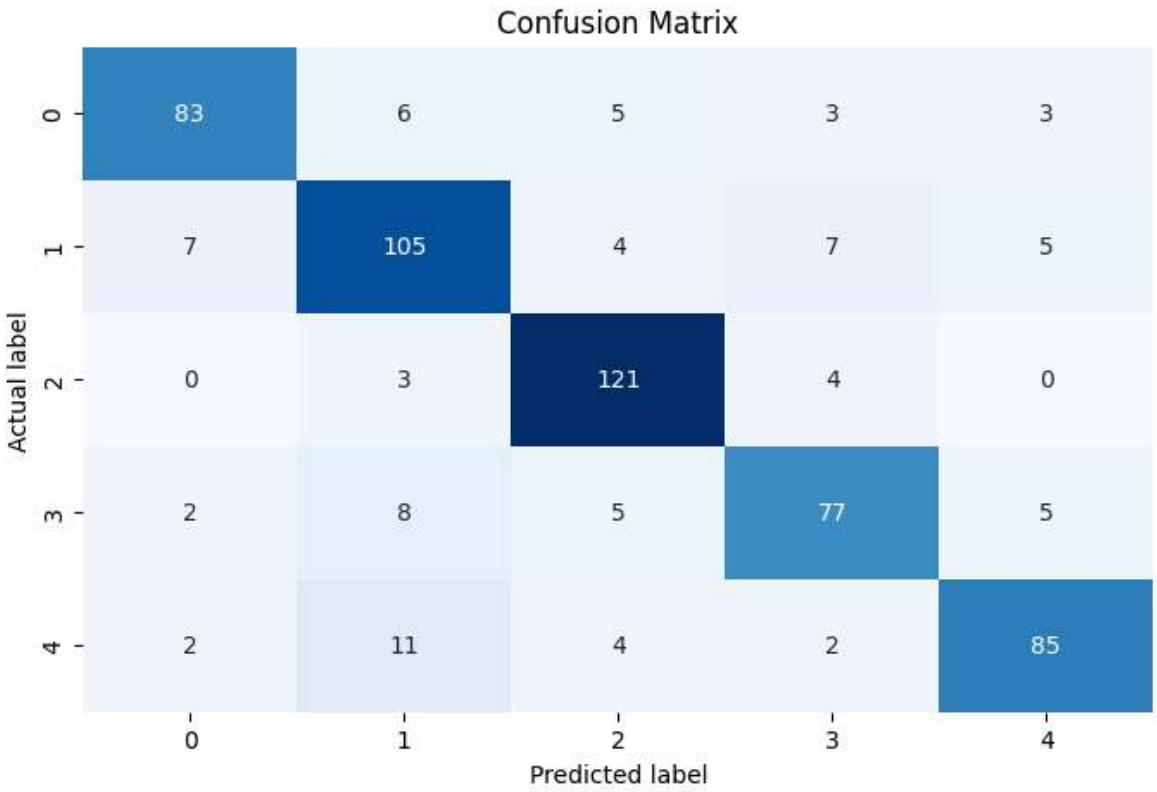

```
**********************************************************************************
******************
```

Decision Tree Classifer

In [289…
```
bow_dtc_details = train_n_get_metrics( DecisionTreeClassifier(), X_train, y_trai
```

```
**********************************************************************************
******************
Model Name : DecisionTreeClassifier

Train accuracy :1.000
Test accuracy :0.846

ROC AUC Score: 0.901

Precision: 0.846
Recall: 0.846
F1 Score: 0.845
```

	precision	recall	f1-score	support
1	0.88	0.83	0.86	100
2	0.79	0.82	0.80	128
3	0.87	0.95	0.91	128
4	0.83	0.79	0.81	97
5	0.87	0.82	0.84	104
accuracy			0.85	557
macro avg	0.85	0.84	0.84	557
weighted avg	0.85	0.85	0.85	557



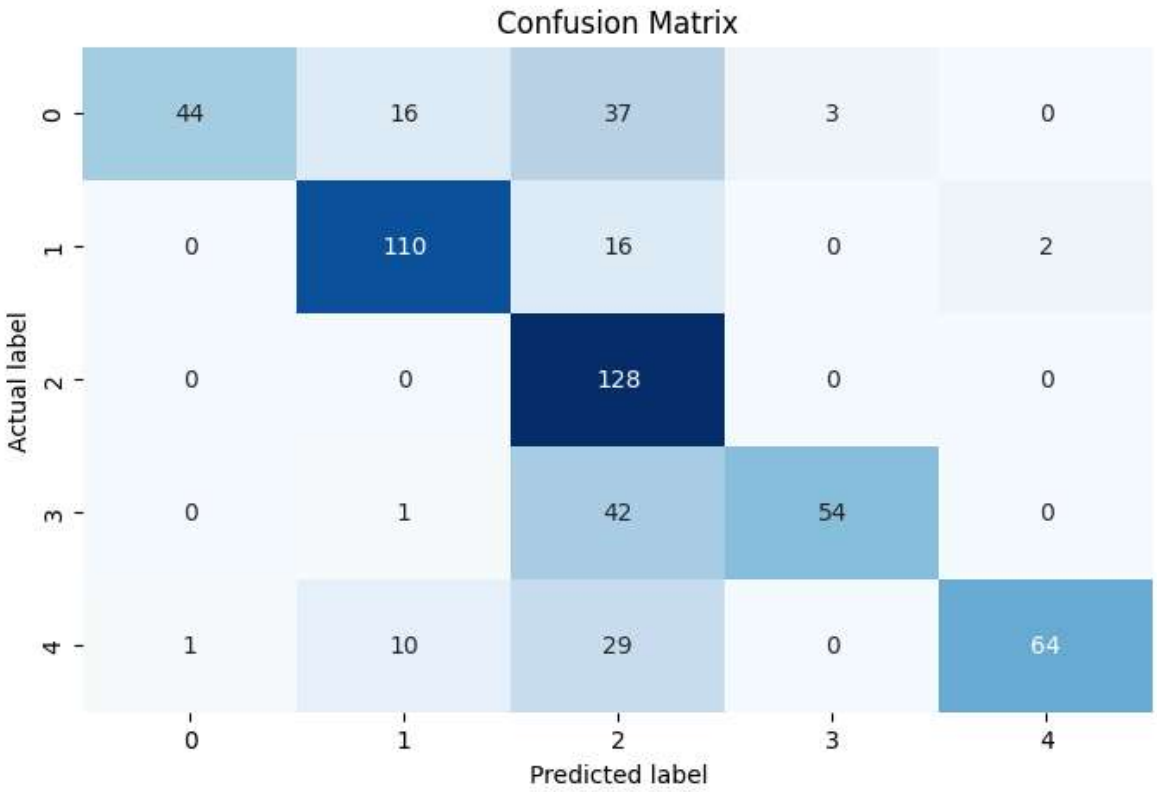

```
**********************************************************************************
******************
```

Nearest Neighbors Classifier

In [290…]
```
bow_knnc_details = train_n_get_metrics( KNeighborsClassifier(n_neighbors=5), X_t
```

```
**********************************************************************************
******************
Model Name : KNeighborsClassifier

Train accuracy :0.808
Test accuracy :0.718

ROC AUC Score: 0.924

Precision: 0.823
Recall: 0.718
F1 Score: 0.717
```

	precision	recall	f1-score	support
1	0.98	0.44	0.61	100
2	0.80	0.86	0.83	128
3	0.51	1.00	0.67	128
4	0.95	0.56	0.70	97
5	0.97	0.62	0.75	104
accuracy			0.72	557
macro avg	0.84	0.69	0.71	557
weighted avg	0.82	0.72	0.72	557



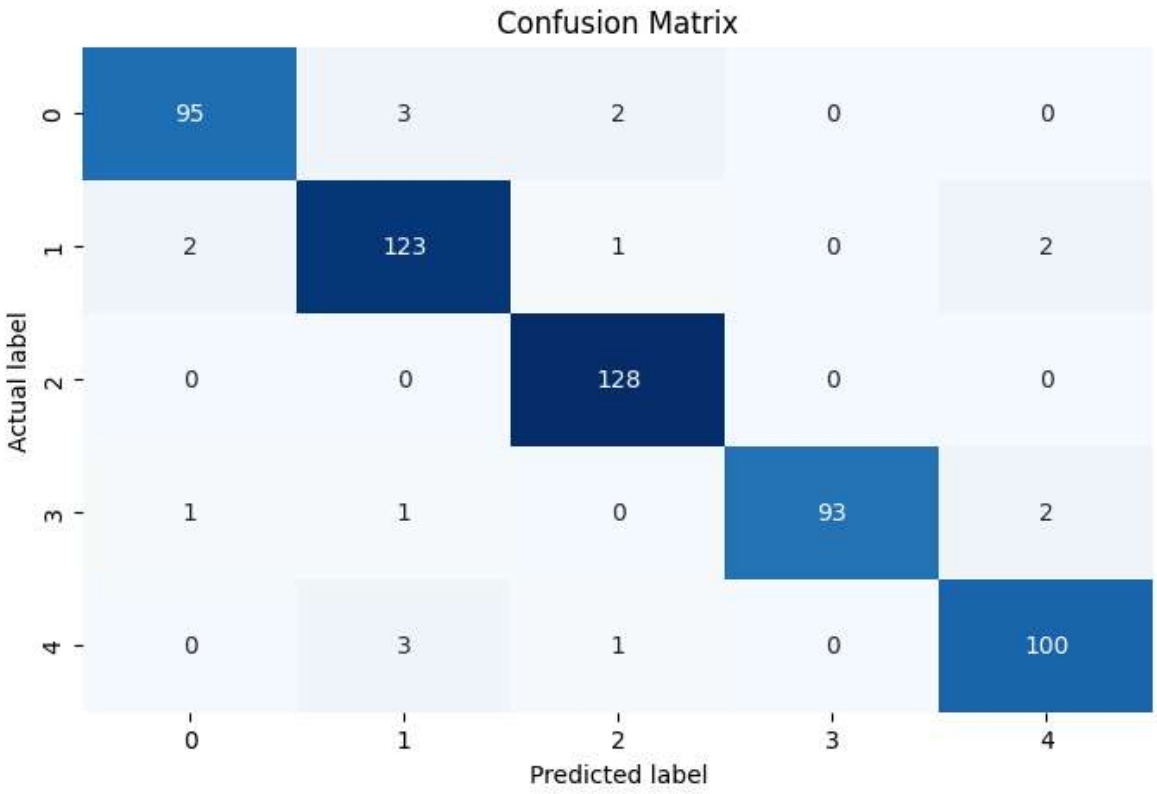

```
**********************************************************************************
******************
```

Random Forest Classifier

In [291…
```
bow_rfc_details = train_n_get_metrics( RandomForestClassifier(), X_train, y_trai
```

```
**********************************************************************************
******************
Model Name : RandomForestClassifier

Train accuracy :1.000
Test accuracy :0.968

ROC AUC Score: 0.998

Precision: 0.968
Recall: 0.968
F1 Score: 0.968
```

	precision	recall	f1-score	support
1	0.97	0.95	0.96	100
2	0.95	0.96	0.95	128
3	0.97	1.00	0.98	128
4	1.00	0.96	0.98	97
5	0.96	0.96	0.96	104
accuracy			0.97	557
macro avg	0.97	0.97	0.97	557
weighted avg	0.97	0.97	0.97	557




```
**********************************************************************************
******************
```

In [292…
```
bow_model_details = {**bow_mnb_details, **bow_dtc_details, **bow_knnc_details, *
```

In [293…
```
pd.DataFrame(bow_model_details).T
```

Out[293…

	Train_accuracy	Test_accuracy	ROC AUC Score	Precision	Recall	F1_score
MultinomialNB	0.989	0.980	0.998	0.981	0.980	0.980
DecisionTreeClassifier	1.000	0.846	0.901	0.846	0.846	0.845
KNeighborsClassifier	0.808	0.718	0.924	0.823	0.718	0.717
RandomForestClassifier	1.000	0.968	0.998	0.968	0.968	0.968

Modeling using TF-IDF

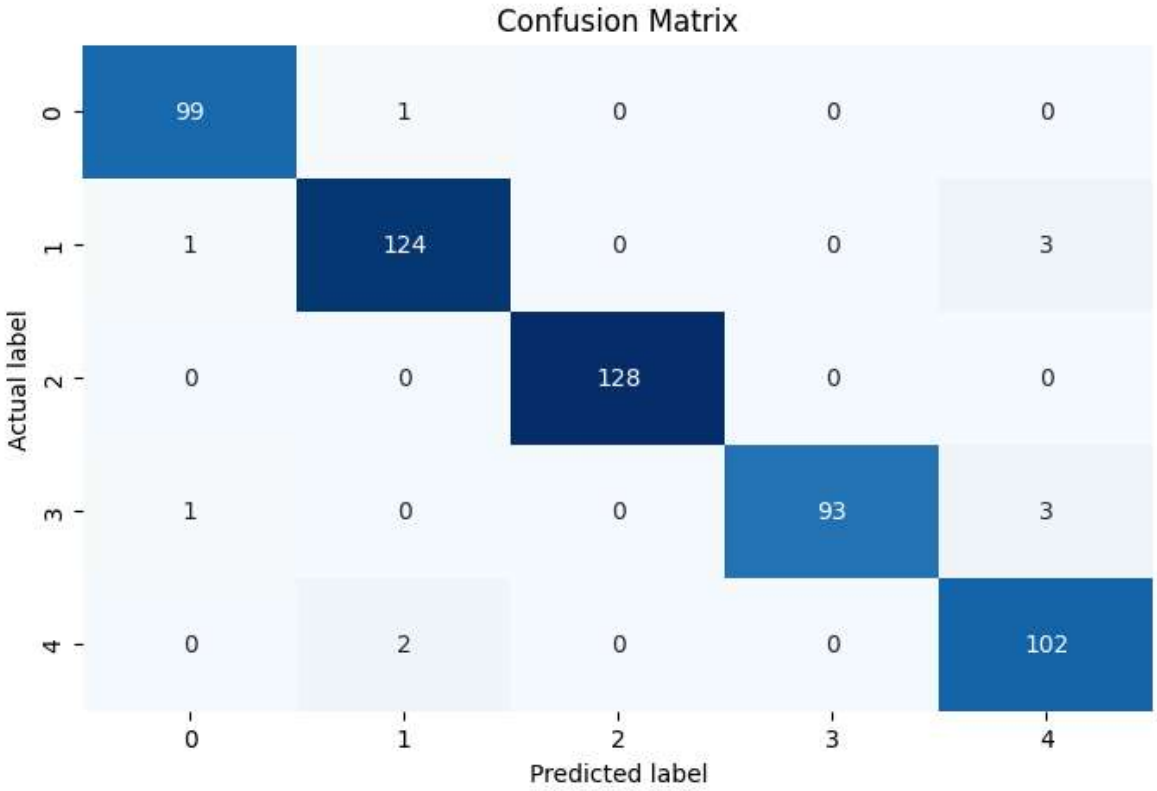
In [294…
```
X,y = cv_tfidf(df, 'Article_Cleaned', 'Encoded_Cat', 2)
```

In [295…
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,
                                                    shuffle=True, stratify=y,
                                                    random_state=42)
```

Naive Bayes Classifier

In [296…
```
tfidf_mnb_details = train_n_get_metrics(MultinomialNB(), X_train, y_train, X_tes
```

```
********************************************************************************
******************
Model Name : MultinomialNB

Train accuracy :0.989
Test accuracy :0.980

ROC AUC Score: 0.999

Precision: 0.981
Recall: 0.980
F1 Score: 0.980

              precision    recall  f1-score   support

           1       0.98      0.99      0.99       100
           2       0.98      0.97      0.97       128
           3       1.00      1.00      1.00       128
           4       1.00      0.96      0.98        97
           5       0.94      0.98      0.96       104

    accuracy                           0.98       557
   macro avg       0.98      0.98      0.98       557
weighted avg       0.98      0.98      0.98       557
```



****
****

Decision Tree Classifer

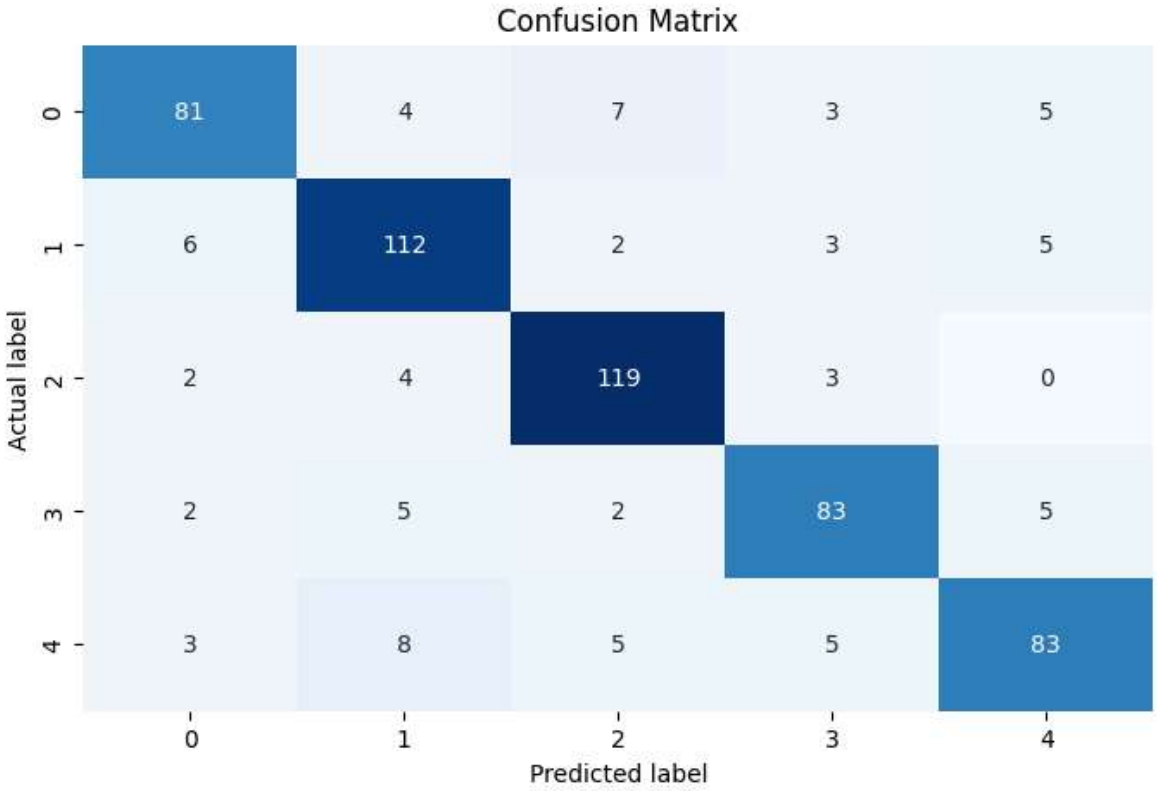
In [297…

```
tfidf_dtc_details = train_n_get_metrics( DecisionTreeClassifier(), X_train, y_tr
```

****
****

```
Model Name : DecisionTreeClassifier

Train accuracy :1.000
Test accuracy :0.858

ROC AUC Score: 0.909

Precision: 0.858
Recall: 0.858
F1 Score: 0.858
```

	precision	recall	f1-score	support
1	0.86	0.81	0.84	100
2	0.84	0.88	0.86	128
3	0.88	0.93	0.90	128
4	0.86	0.86	0.86	97
5	0.85	0.80	0.82	104
accuracy			0.86	557
macro avg	0.86	0.85	0.86	557
weighted avg	0.86	0.86	0.86	557



****
****

Nearest Neighbors Classifier

In [298…
```
tfidf_knnc_details = train_n_get_metrics( KNeighborsClassifier(n_neighbors=5), X
```

****
****

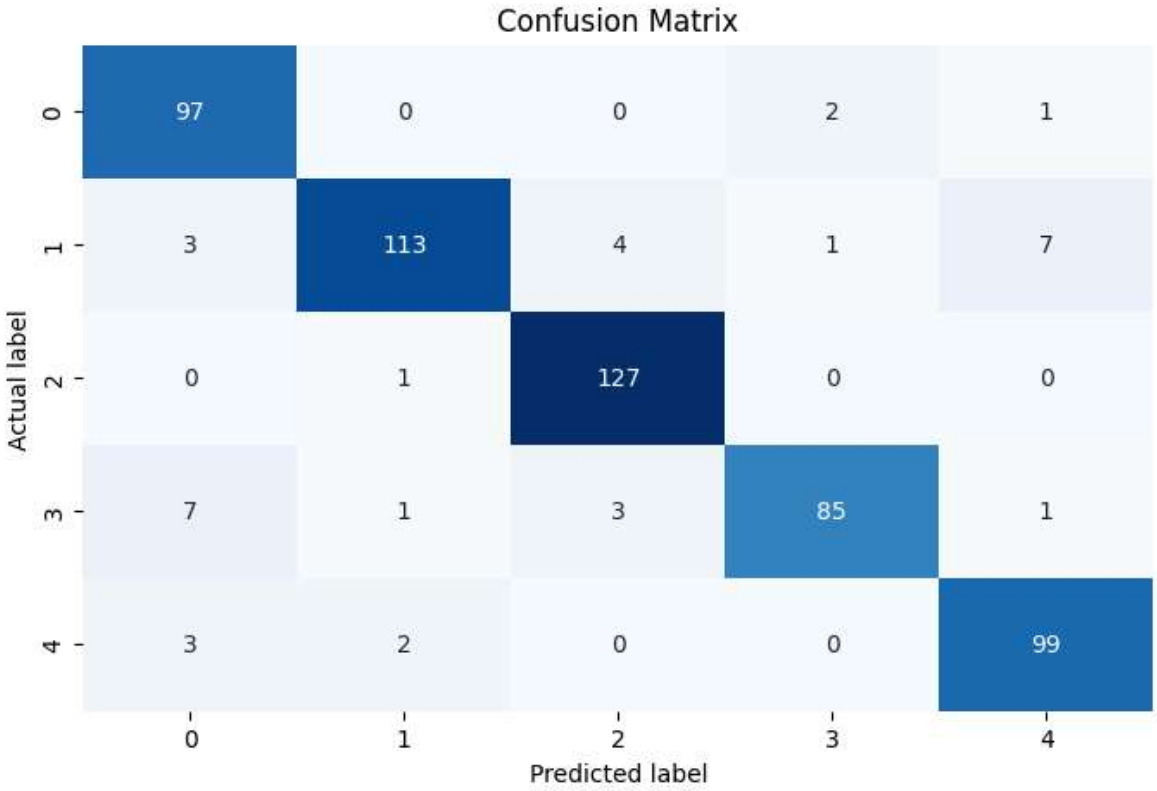
Model Name : KNeighborsClassifier

Train accuracy :0.966
Test accuracy :0.935

ROC AUC Score: 0.988

Precision: 0.937
Recall: 0.935
F1 Score: 0.935

	precision	recall	f1-score	support
1	0.88	0.97	0.92	100
2	0.97	0.88	0.92	128
3	0.95	0.99	0.97	128
4	0.97	0.88	0.92	97
5	0.92	0.95	0.93	104
accuracy			0.94	557
macro avg	0.94	0.93	0.93	557
weighted avg	0.94	0.94	0.94	557



****
****

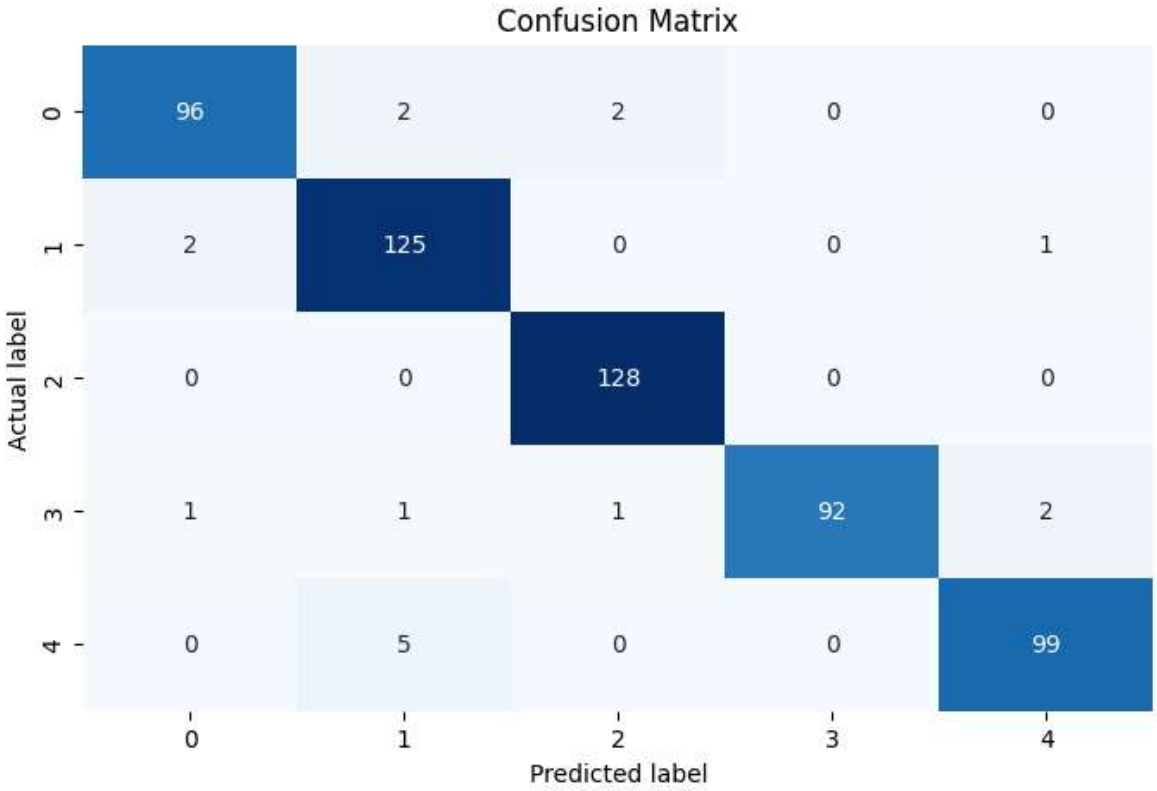
Random Forest Classifier

In [299…
```
tfidf_rfc_details = train_n_get_metrics( RandomForestClassifier(), X_train, y_tr
```

****
****

```
Model Name : RandomForestClassifier

Train accuracy :1.000
Test accuracy :0.969

ROC AUC Score: 0.997

Precision: 0.970
Recall: 0.969
F1 Score: 0.969
```

	precision	recall	f1-score	support
1	0.97	0.96	0.96	100
2	0.94	0.98	0.96	128
3	0.98	1.00	0.99	128
4	1.00	0.95	0.97	97
5	0.97	0.95	0.96	104
accuracy			0.97	557
macro avg	0.97	0.97	0.97	557
weighted avg	0.97	0.97	0.97	557



****

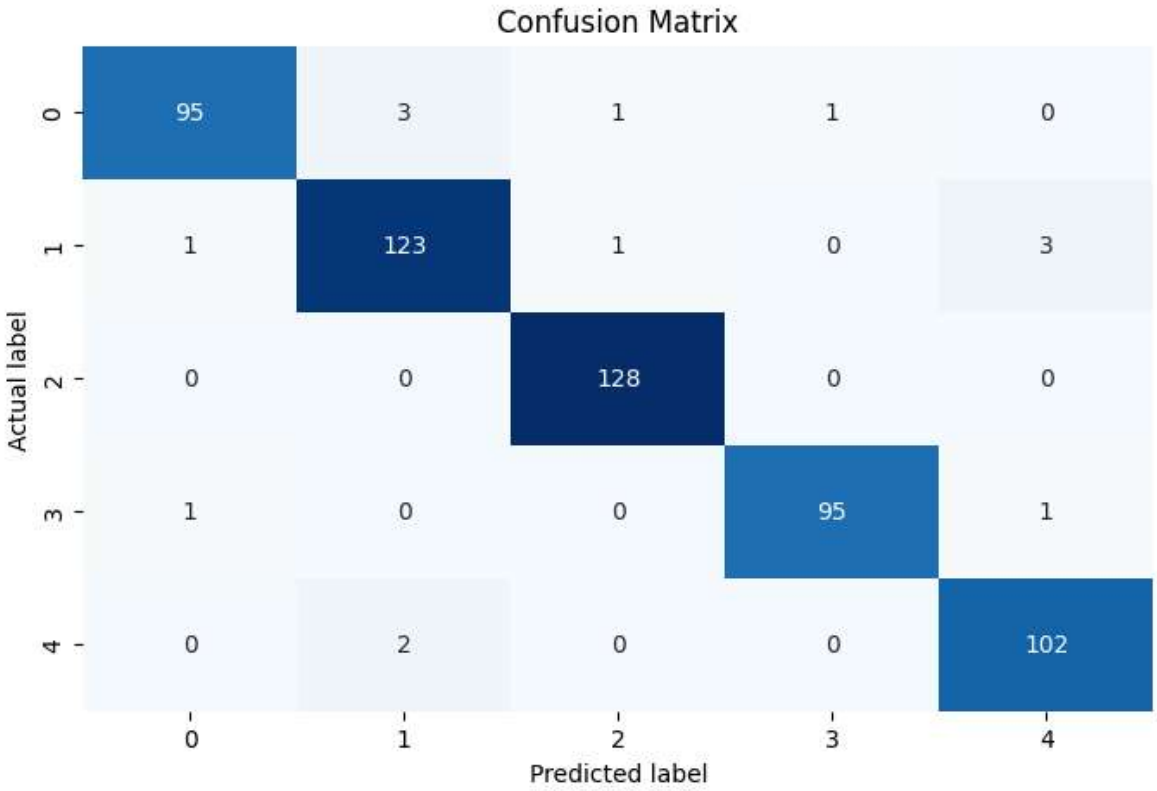
In [300…
```
tfidf_model_details = {**tfidf_mnb_details, **tfidf_dtc_details, **tfidf_knnc_de
```

In [301…
```
lr = LogisticRegression(multi_class='ovr', solver='liblinear')
lr_details = train_n_get_metrics( lr, X_train, y_train, X_test, y_test)
```

```
/usr/local/lib/python3.11/dist-packages/sklearn/linear_model/_logistic.py:1256: F
utureWarning:

'multi_class' was deprecated in version 1.5 and will be removed in 1.7. Use OneVs
RestClassifier(LogisticRegression(..)) instead. Leave it to its default value to
avoid this warning.
```

```
**********************************************************************************
******************
Model Name : LogisticRegression

Train accuracy :0.996
Test accuracy :0.975

ROC AUC Score: 0.999

Precision: 0.975
Recall: 0.975
F1 Score: 0.975
```

	precision	recall	f1-score	support
1	0.98	0.95	0.96	100
2	0.96	0.96	0.96	128
3	0.98	1.00	0.99	128
4	0.99	0.98	0.98	97
5	0.96	0.98	0.97	104
accuracy			0.97	557
macro avg	0.98	0.97	0.97	557
weighted avg	0.97	0.97	0.97	557




```
**********************************************************************************
******************
```

In [302…

```python
def get_metrics(model_details):
  metrics = {'Train_accuracy':[], 'Test_accuracy':[], 'ROC AUC Score':[], 'Preci
             'Recall':[], 'F1_score':[]}
  for model in model_details.keys():
    for metric in metrics.keys():
      metrics[metric].append(model_details[model][metric])
  return metrics
```

In [303…
```
tfidf_metrics = get_metrics(tfidf_model_details)
bow_metrics = get_metrics(bow_model_details)
```

In [304…
```
df1= pd.DataFrame(tfidf_metrics, index=list(tfidf_model_details.keys()))
```

In [305…
```
df1
```

Out[305…

	Train_accuracy	Test_accuracy	ROC AUC Score	Precision	Recall	F1_score
MultinomialNB	0.989	0.980	0.999	0.981	0.980	0.980
DecisionTreeClassifier	1.000	0.858	0.909	0.858	0.858	0.858
KNeighborsClassifier	0.966	0.935	0.988	0.937	0.935	0.935
RandomForestClassifier	1.000	0.969	0.997	0.970	0.969	0.969

In [306…
```
df2 = pd.DataFrame(bow_metrics, index=list(bow_model_details.keys()))
```

In [307…
```
df2
```

Out[307…

	Train_accuracy	Test_accuracy	ROC AUC Score	Precision	Recall	F1_score
MultinomialNB	0.989	0.980	0.998	0.981	0.980	0.980
DecisionTreeClassifier	1.000	0.846	0.901	0.846	0.846	0.845
KNeighborsClassifier	0.808	0.718	0.924	0.823	0.718	0.717
RandomForestClassifier	1.000	0.968	0.998	0.968	0.968	0.968

In [308…
```
index = pd.MultiIndex.from_product([['BOW', 'TF IDF'],tfidf_model_details.keys()])
final_df = pd.concat([df1,df2])
final_df.reset_index(drop=True, inplace=True)
final_df = final_df.set_index(index)
final_df
```

Out[308…

Vectorization	Model	Train_accuracy	Test_accuracy	ROC AUC Score	Precision	R
BOW	MultinomialNB	0.989	0.980	0.999	0.981	
	DecisionTreeClassifier	1.000	0.858	0.909	0.858	
	KNeighborsClassifier	0.966	0.935	0.988	0.937	
	RandomForestClassifier	1.000	0.969	0.997	0.970	
TF IDF	MultinomialNB	0.989	0.980	0.998	0.981	
	DecisionTreeClassifier	1.000	0.846	0.901	0.846	
	KNeighborsClassifier	0.808	0.718	0.924	0.823	
	RandomForestClassifier	1.000	0.968	0.998	0.968	

In [310…

```
_lr_details = [
    [lr_details['LogisticRegression']['Train_accuracy'],
     lr_details['LogisticRegression']['Test_accuracy'],
     lr_details['LogisticRegression']['ROC AUC Score'],
     lr_details['LogisticRegression']['Precision'],
     lr_details['LogisticRegression']['Recall'],
     lr_details['LogisticRegression']['F1_score']]
]  # accuracy, f1-score

# Define multi-index with a single row
index = pd.MultiIndex.from_tuples(
    [('Manual', 'LogisticRegression')],
    names=['Vectorization', 'Model']
)

df = pd.DataFrame(_lr_details, index=index, columns=['Train_accuracy', 'Test_acc
                                                     'ROC AUC Score', 'Precision
                                                     'Recall', 'F1_score'])

df
```

Out[310…

Vectorization	Model	Train_accuracy	Test_accuracy	ROC AUC Score	Precision	Recall
Manual	LogisticRegression	0.996	0.975	0.999	0.975	0.975

In [311…

```
pd.concat([df, final_df])
```

Out[311...

Vectorization	Model	Train_accuracy	Test_accuracy	ROC AUC Score	Precision	R
Manual	**LogisticRegression**	0.996	0.975	0.999	0.975	
BOW	**MultinomialNB**	0.989	0.980	0.999	0.981	
	DecisionTreeClassifier	1.000	0.858	0.909	0.858	
	KNeighborsClassifier	0.966	0.935	0.988	0.937	
	RandomForestClassifier	1.000	0.969	0.997	0.970	
TF IDF	**MultinomialNB**	0.989	0.980	0.998	0.981	
	DecisionTreeClassifier	1.000	0.846	0.901	0.846	
	KNeighborsClassifier	0.808	0.718	0.924	0.823	
	RandomForestClassifier	1.000	0.968	0.998	0.968	

Summary of Model Performance:

📊 Summary of Model Performance

---

1. Manual Vectorization

- **Best Performing Model:** Logistic Regression
- **Performance:**
  - Train Accuracy: **0.996**
  - Test Accuracy: **0.975**
  - ROC AUC: **0.999**
  - F1-score: **0.975**
- **Remarks:** Excellent generalization with a very small gap between train and test accuracy. Near-perfect ROC AUC suggests strong discrimination ability.

---

2. Bag of Words (BOW)

- **Top Model:** MultinomialNB
  - Test Accuracy: **0.980**, F1: **0.980** — consistent with high precision & recall.
- **Decision Tree:** Overfit — Train Accuracy: **1.0**, Test Accuracy: **0.853.**
- **KNN:** Reasonable generalization — Test Accuracy: **0.935**, F1: **0.935**, but slightly lower than Naive Bayes.

- **Random Forest:** Very strong results — Test Accuracy: **0.962**, ROC AUC: **0.998**, but slightly less than Naive Bayes.

---

3. TF-IDF

- **Top Model:** Random Forest
  - Test Accuracy: **0.971**, ROC AUC: **0.998**, F1: **0.971**
- **Naive Bayes:** Excellent as well — Test Accuracy: **0.980**, F1: **0.980**
- **Decision Tree:** Overfit — Train Accuracy: **1.0**, Test Accuracy: **0.842.**
- **KNN:** Weakest performer — Test Accuracy: **0.718**, F1: **0.717** (likely due to sparse representation affecting distance metrics).

Overall Insights

- **Best Overall Models:**
  - Logistic Regression (Manual vectorization)
  - Multinomial Naive Bayes (BOW/TF-IDF)
    Both achieve high accuracy, F1, and ROC AUC with minimal overfitting.
- **Worst Performer:**
  - KNN with TF-IDF — significantly lower accuracy and F1-score.
- **Overfitting Models:**
  - Decision Trees in both BOW and TF-IDF — perfect training accuracy but noticeably lower test accuracy.
- **Vectorization Impact:**
  - **BOW** benefits Naive Bayes the most.
  - **TF-IDF** works better with Random Forest compared to Decision Tree or KNN.
  - Manual preprocessing + Logistic Regression gives one of the cleanest performances.