Barutcu_Semih_BBC_News

March 30, 2021

1 BBC News NLP Project

In this study, I used the BBC news dataset. The dataset was produced for the shared publication.

- D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.

There are 5 different categories of news(business, entertainment, politics, sport, tech) and they are consisted of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005.

My first goal was identifying G20 countries in the news and answering total count related questions. Secondly, I focused to find themes of each categories by using Latent Dirichlet Allocation (LDA).

[1]:
```
import gensim
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import STOPWORDS
from nltk.stem import WordNetLemmatizer, SnowballStemmer
from nltk.stem.porter import *
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

[2]:
```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\sbaru\AppData\Roaming\nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\sbaru\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\sbaru\AppData\Roaming\nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[2]: True

1.1 1) Loading the Dataset

[3]:
```
# Reading texts by per folder and relative lengths.
# They recorded the 'text' list by the category and the body.
# Then converted to a dataframe as 'bbc'

files_list = [['business', 511], ['entertainment', 387], ['politics', 418],
              ['sport', 512], ['tech', 402]]
text = []

for file, length in files_list:
    for i in np.arange(1,10):
        filename = file + '\\00'+str(i)+'.txt'
        with open(filename) as f:
            lines = f.readlines()
            lines = ' '.join([line.strip() for line in lines])
            text.append([file, lines])

    for i in np.arange(10,100):
        filename = file + '\\0'+str(i)+'.txt'
        with open(filename) as f:
            lines = f.readlines()
            lines = ' '.join([line.strip() for line in lines])
            text.append([file, lines])

    for i in np.arange(100,length):
        filename = file + '\\'+str(i)+'.txt'
        with open(filename) as f:
            lines = f.readlines()
            lines = ' '.join([line.strip() for line in lines])
            text.append([file, lines])

bbc = pd.DataFrame(text, columns=['category', 'text'])
bbc['category'] = pd.Categorical(bbc['category'])
```

1.2 2) Preparing the Data

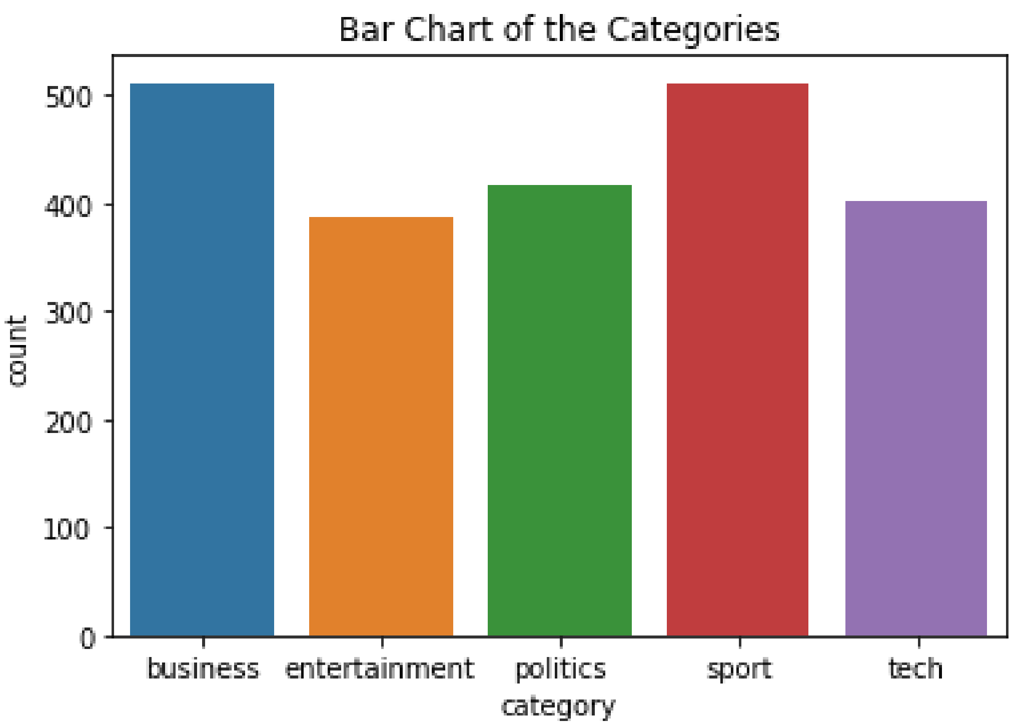
[4]:
```
# Resource code:
# Remove punctuation

bbc['text'] = bbc['text'].map(lambda x: re.sub('[,\.!?]', '', x))
# Convert the titles to lowercase
bbc['text'] = bbc['text'].map(lambda x: re.sub('\n', '', x))
bbc['text'] = bbc['text'].map(lambda x: x.lower())
```

```
# Print out the first rows of papers
bbc['text'].head()
```

[4]:
```
0    ad sales boost time warner profit  quarterly p…
1    dollar gains on greenspan speech  the dollar h…
2    yukos unit buyer faces loan claim  the owners …
3    high fuel prices hit ba's profits  british air…
4    pernod takeover talk lifts domecq  shares in u…
Name: text, dtype: object
```

1.3 3) Exploratary Data Analysis

[5]:
```
sns.countplot(x=bbc.category)
plt.title('Bar Chart of the Categories');
```



[6]:
```
# Import the wordcloud library
from wordcloud import WordCloud
# Join the different processed titles together.
long_string = ','.join(list(bbc['text'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,␣
 ↪contour_width=3, contour_color='steelblue')
# Generate a word cloud
```

```
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

[6]:

1.3.1 Assumptions

After a google search, I learned that G20 includes not only countries but also the European Union. I only added the most common abbreviations for the United Kingdom, the United States of America, and the European Union. Because 'us' is a common word in English I looked for 'the us'. Also, I observe that 'usa' is used frequently and consider it in my keywords. Other than that, I didn't add lots of possible keywords such as countries and states in the unions, nationalities, languages, and cities which are obviously identified by certain countries but I preferred to keep the scope more restricted.

1.4 Part 1

[7]:

```
# G20 countries with the most common abbrevations

G_20 = ['argentina', 'australia', 'brazil', 'canada', 'china', 'france', 'germany', 'japan',
        'india', 'indonesia', 'italy', 'mexico', 'russia', 'south africa', 'saudi arabia',
        'south korea', 'turkey', 'united kingdom', 'united states', 'european union',
        'uk', 'eu', 'the us', 'usa']
```

[8]:

```
# Adding columns for every G20 countries respectively their how many they are mentioned in every news
bbc2 = bbc.copy()
```

```python
for country in G_20:
    f = lambda x: x.text.count(country)
    bbc2[country] = bbc2.apply(f, axis = 'columns')

# The first 10 observations with newly added columns
bbc2.head().T
```

[8]:
```
                                                                 0  \
category                                                  business
text              ad sales boost time warner profit  quarterly p…
argentina                                                        0
australia                                                        0
brazil                                                           0
canada                                                           0
china                                                            0
france                                                           0
germany                                                          0
japan                                                            0
india                                                            0
indonesia                                                        0
italy                                                            0
mexico                                                           0
russia                                                           0
south africa                                                     0
saudi arabia                                                     0
south korea                                                      0
turkey                                                           0
united kingdom                                                   0
united states                                                    0
european union                                                   0
uk                                                               0
eu                                                               2
the us                                                           1
usa                                                              0

                                                                 1  \
category                                                  business
text              dollar gains on greenspan speech  the dollar h…
argentina                                                        0
australia                                                        0
brazil                                                           0
canada                                                           0
china                                                            2
france                                                           0
germany                                                          0
japan                                                            0
```

```
india                                                              0
indonesia                                                          0
italy                                                              0
mexico                                                             0
russia                                                             0
south africa                                                       0
saudi arabia                                                       0
south korea                                                        0
turkey                                                             0
united kingdom                                                     0
united states                                                      0
european union                                                     0
uk                                                                 0
eu                                                                 3
the us                                                             5
usa                                                                0

                                                                   2  \
category                                                    business
text            yukos unit buyer faces loan claim  the owners …
argentina                                                          0
australia                                                          0
brazil                                                             0
canada                                                             0
china                                                              0
france                                                             0
germany                                                            0
japan                                                              0
india                                                              0
indonesia                                                          0
italy                                                              0
mexico                                                             0
russia                                                             2
south africa                                                       0
saudi arabia                                                       0
south korea                                                        0
turkey                                                             0
united kingdom                                                     0
united states                                                      0
european union                                                     0
uk                                                                 6
eu                                                                 1
the us                                                             0
usa                                                                0

                                                                   3  \
category                                                    business
```

```
text            high fuel prices hit ba's profits  british air…
argentina                                                    0
australia                                                    0
brazil                                                       0
canada                                                       0
china                                                        0
france                                                       0
germany                                                      0
japan                                                        0
india                                                        0
indonesia                                                    0
italy                                                        0
mexico                                                       0
russia                                                       0
south africa                                                 0
saudi arabia                                                 0
south korea                                                  0
turkey                                                       0
united kingdom                                               0
united states                                                1
european union                                               0
uk                                                           0
eu                                                           0
the us                                                       0
usa                                                          0

                                                             4
category                                              business
text            pernod takeover talk lifts domecq  shares in u…
argentina                                                    0
australia                                                    0
brazil                                                       0
canada                                                       0
china                                                        0
france                                                       1
germany                                                      0
japan                                                        0
india                                                        0
indonesia                                                    0
italy                                                        0
mexico                                                       0
russia                                                       0
south africa                                                 0
saudi arabia                                                 0
south korea                                                  0
turkey                                                       0
united kingdom                                               0
```

```
united states                                                   0
european union                                                  0
uk                                                              1
eu                                                              3
the us                                                          0
usa                                                             0
```

[9]:
```python
# Combining  the columns which are used for the same monarchy, country, and␣
↪union

double_columns_countries = [['united kingdom', 'uk'], ['united states', 'the␣
↪us'],
                            ['united states', 'usa'], ['european union', 'eu']]

for i, j in double_columns_countries:
    bbc2[i + ' total'] = bbc2[i] + bbc2[j]
```

[10]:
```python
# dropping aggregated columns

bbc2 = bbc2.drop(['united kingdom', 'uk', 'united states', 'the us',
                'united states', 'usa', 'european union', 'eu'], axis=1)
```

1.4.1 Answer 1.1

[11]:
```python
G_20bbc = bbc2.select_dtypes(include='int64')

bbc2['total_count'] = G_20bbc.apply(lambda x: x.sum(), axis='columns')

print("Total number of news that incleded G20 countries: " +␣
↪str(bbc2[bbc2['total_count'] > 0].count()['total_count']))
```

```
Total number of news that incleded G20 countries: 1506
```

1.4.2 Answer 1.2

[12]:
```python
G_20_boolean = G_20bbc > 0

bbc2['total_countries'] = G_20_boolean.apply(lambda x: x.sum(), axis='columns')

print("Total number of news that incleded several G20 countries: " +
      str(bbc2[bbc2['total_countries'] > 1].count()['total_countries']))
```

```
Total number of news that incleded several G20 countries: 703
```

1.5 Part 2

In this part, I focused the themes for every main topic of news. I used LDA topic modeling by using similar code that I shared below. You can find 5 different related themes for every section below.

Resource codes: https://github.com/priya-dwivedi/Deep-Learning/blob/master/topic_modeling/LDA_Newsgrou

[13]:
```
def lemmatize_stemming(text):
    return stemmer.stem(WordNetLemmatizer().lemmatize(text, pos='v'))

# Tokenize and lemmatize
def preprocess(text):
    result=[]
    for token in gensim.utils.simple_preprocess(text) :
        if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) >␣
 ↪3:
            result.append(lemmatize_stemming(token))

    return result
```

[14]:
```
category_list = ['business', 'entertainment', 'politics', 'sport', 'tech']
lda_model_list = []
stemmer = SnowballStemmer("english")

for category in category_list:
    processed_docs = []

    for doc in bbc.loc[bbc['category'] == category]['text']:
        processed_docs.append(preprocess(doc))

    dictionary = gensim.corpora.Dictionary(processed_docs)

    '''
    Remove very rare and very common words:

    - words appearing less than 15 times
    - words appearing in more than 10% of all documents
    '''
    dictionary.filter_extremes(no_below=15, no_above=0.1, keep_n= 100000)

    bow_corpus = [dictionary.doc2bow(doc) for doc in processed_docs]

    lda_model =  gensim.models.LdaMulticore(bow_corpus,
                                   num_topics = 5,
                                   id2word = dictionary,
                                   passes = 10,
                                   workers = 2)

    lda_model_list.append(lda_model)
```

[15]:

```
counter = 0
for category in category_list:

    model = lda_model_list[counter]
    for idx, topic in model.print_topics(-1):
        print(category.upper() + ':')
        print("Topic: {} \nWords: {}".format(idx, topic ))
        print("\n")

    counter += 1
```

```
BUSINESS:
Topic: 0
Words: 0.055*"yuko" + 0.034*"india" + 0.031*"russian" + 0.027*"court" +
0.025*"russia" + 0.020*"gazprom" + 0.016*"rosneft" + 0.016*"bankruptci" +
0.015*"auction" + 0.014*"indian"


BUSINESS:
Topic: 1
Words: 0.035*"airlin" + 0.023*"insur" + 0.014*"commiss" + 0.012*"investig" +
0.012*"damag" + 0.011*"fuel" + 0.009*"disast" + 0.009*"travel" + 0.008*"affect"
+ 0.008*"asia"


BUSINESS:
Topic: 2
Words: 0.019*"deficit" + 0.012*"japan" + 0.011*"worldcom" + 0.010*"index" +
0.010*"telecom" + 0.010*"fraud" + 0.010*"currenc" + 0.009*"elect" + 0.008*"bush"
+ 0.007*"manufactur"


BUSINESS:
Topic: 3
Words: 0.019*"retail" + 0.019*"club" + 0.017*"deutsch" + 0.012*"german" +
0.010*"unemploy" + 0.009*"card" + 0.009*"mortgag" + 0.009*"board" +
0.009*"christma" + 0.009*"takeov"


BUSINESS:
Topic: 4
Words: 0.013*"list" + 0.011*"project" + 0.009*"contract" + 0.009*"brand" +
0.008*"worker" + 0.008*"maker" + 0.008*"stake" + 0.007*"factori" + 0.007*"centr"
+ 0.007*"propos"
```

```
ENTERTAINMENT:
Topic: 0
Words: 0.016*"radio" + 0.014*"danc" + 0.012*"elvi" + 0.011*"richard" +
0.009*"concert" + 0.009*"opera" + 0.009*"saturday" + 0.009*"histori" +
0.008*"boy" + 0.008*"mari"


ENTERTAINMENT:
Topic: 1
Words: 0.014*"rapper" + 0.010*"franz" + 0.010*"concert" + 0.010*"citi" +
0.010*"ferdinand" + 0.010*"ticket" + 0.010*"rais" + 0.010*"hous" +
0.009*"brother" + 0.009*"contest"


ENTERTAINMENT:
Topic: 2
Words: 0.013*"stone" + 0.012*"black" + 0.011*"soul" + 0.011*"list" +
0.010*"robbi" + 0.009*"brit" + 0.009*"tour" + 0.009*"britain" + 0.008*"episod" +
0.008*"william"


ENTERTAINMENT:
Topic: 3
Words: 0.023*"foxx" + 0.022*"babi" + 0.019*"dollar" + 0.019*"jackson" +
0.019*"vera" + 0.018*"drake" + 0.016*"bafta" + 0.016*"jami" + 0.016*"eastwood" +
0.016*"scorses"


ENTERTAINMENT:
Topic: 4
Words: 0.013*"christma" + 0.013*"court" + 0.012*"documentari" + 0.011*"weekend"
+ 0.010*"claim" + 0.010*"meet" + 0.010*"action" + 0.010*"pictur" +
0.009*"french" + 0.009*"anim"


POLITICS:
Topic: 0
Words: 0.034*"hunt" + 0.026*"blunkett" + 0.022*"trial" + 0.021*"clark" +
0.021*"suspect" + 0.017*"arrest" + 0.016*"judg" + 0.015*"prison" +
0.014*"terrorist" + 0.011*"inquiri"


POLITICS:
Topic: 1
Words: 0.015*"card" + 0.014*"cut" + 0.011*"answer" + 0.011*"candid" +
0.010*"young" + 0.009*"advic" + 0.009*"milburn" + 0.008*"elector" +
0.008*"societi" + 0.008*"wast"
```

```
POLITICS:
Topic: 2
Words: 0.017*"minimum" + 0.016*"busi" + 0.015*"sentenc" + 0.013*"wait" +
0.012*"muslim" + 0.012*"campbel" + 0.012*"job" + 0.012*"pay" + 0.011*"inform" +
0.010*"employ"


POLITICS:
Topic: 3
Words: 0.022*"asylum" + 0.020*"women" + 0.017*"straw" + 0.012*"book" +
0.010*"constitut" + 0.009*"europ" + 0.008*"peac" + 0.008*"visit" + 0.008*"minor"
+ 0.006*"feel"


POLITICS:
Topic: 4
Words: 0.015*"univers" + 0.014*"student" + 0.014*"scottish" + 0.013*"scotland" +
0.012*"market" + 0.012*"poster" + 0.010*"debt" + 0.009*"research" +
0.008*"financ" + 0.008*"duti"


SPORT:
Topic: 0
Words: 0.019*"indoor" + 0.016*"zealand" + 0.014*"lion" + 0.013*"britain" +
0.012*"holm" + 0.012*"johnson" + 0.011*"marathon" + 0.010*"tour" +
0.010*"compet" + 0.010*"gold"


SPORT:
Topic: 1
Words: 0.033*"robinson" + 0.018*"seed" + 0.013*"bath" + 0.012*"irish" +
0.012*"leicest" + 0.012*"sullivan" + 0.011*"wasp" + 0.011*"wilkinson" +
0.009*"sale" + 0.008*"dublin"


SPORT:
Topic: 2
Words: 0.020*"penalti" + 0.014*"jone" + 0.011*"shoot" + 0.011*"gara" +
0.009*"henson" + 0.009*"thoma" + 0.008*"yard" + 0.008*"corner" + 0.008*"wide" +
0.007*"edinburgh"


SPORT:
Topic: 3
Words: 0.026*"liverpool" + 0.021*"roddick" + 0.015*"gerrard" + 0.013*"tenni" +
0.013*"deal" + 0.011*"real" + 0.011*"serv" + 0.011*"hewitt" + 0.010*"davi" +
0.010*"feder"
```

SPORT:
Topic: 4
Words: 0.018*"drug" + 0.014*"mourinho" + 0.013*"ferguson" + 0.013*"kenteri" + 0.013*"iaaf" + 0.012*"dope" + 0.011*"wenger" + 0.011*"greek" + 0.011*"thanou" + 0.010*"refere"

TECH:
Topic: 0
Words: 0.016*"xbox" + 0.011*"learn" + 0.010*"china" + 0.008*"team" + 0.007*"handset" + 0.007*"studi" + 0.007*"linux" + 0.007*"speech" + 0.006*"trend" + 0.006*"biggest"

TECH:
Topic: 1
Words: 0.040*"search" + 0.016*"googl" + 0.016*"broadcast" + 0.011*"yahoo" + 0.008*"handset" + 0.008*"listen" + 0.008*"multimedia" + 0.008*"desktop" + 0.007*"voic" + 0.007*"channel"

TECH:
Topic: 2
Words: 0.035*"blog" + 0.024*"attack" + 0.017*"spywar" + 0.012*"infect" + 0.012*"malici" + 0.010*"crimin" + 0.009*"survey" + 0.008*"address" + 0.008*"govern" + 0.008*"spread"

TECH:
Topic: 3
Words: 0.023*"spam" + 0.017*"nintendo" + 0.015*"campaign" + 0.013*"peer" + 0.011*"browser" + 0.011*"handheld" + 0.009*"spammer" + 0.009*"bank" + 0.009*"sourc" + 0.009*"server"

TECH:
Topic: 4
Words: 0.016*"chip" + 0.014*"laptop" + 0.012*"ipod" + 0.011*"mini" + 0.011*"graphic" + 0.009*"intel" + 0.009*"processor" + 0.009*"best" + 0.008*"light" + 0.008*"creativ"