Análise Exploratória

April 7, 2020

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from collections import Counter
import nltk
from wordcloud import WordCloud, STOPWORDS
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.manifold import TSNE
from sklearn.preprocessing import LabelEncoder

%matplotlib inline
```

[2]:
```
df = pd.read_json('../../data/News_Category_Dataset_v2.json', lines=True)
```

[3]:
```
df.describe()
```

[3]:

	category	headline	authors	link	short_description	date
count	200853	200853	200853	200853	200853	200853
unique	41	199344	27993	200812	178353	2309
top	POLITICS	Sunday Roundup		https://www.huffingtonpost.comhttp://recode.ne…		2013-01-17 00:00:00
freq	32739	90	36620	2	19712	
first	NaN	NaN	NaN	NaN	NaN	
last	NaN	NaN	NaN	NaN	NaN	

```
freq                    100
first   2012-01-28 00:00:00
last    2018-05-26 00:00:00
```

[4]:
```
len(df['category'].unique())
```

[4]:
```
41
```

[5]:
```
Counter(df['category']).most_common()
```

[5]:
```
[('POLITICS', 32739),
 ('WELLNESS', 17827),
 ('ENTERTAINMENT', 16058),
 ('TRAVEL', 9887),
 ('STYLE & BEAUTY', 9649),
 ('PARENTING', 8677),
 ('HEALTHY LIVING', 6694),
 ('QUEER VOICES', 6314),
 ('FOOD & DRINK', 6226),
 ('BUSINESS', 5937),
 ('COMEDY', 5175),
 ('SPORTS', 4884),
 ('BLACK VOICES', 4528),
 ('HOME & LIVING', 4195),
 ('PARENTS', 3955),
 ('THE WORLDPOST', 3664),
 ('WEDDINGS', 3651),
 ('WOMEN', 3490),
 ('IMPACT', 3459),
 ('DIVORCE', 3426),
 ('CRIME', 3405),
 ('MEDIA', 2815),
 ('WEIRD NEWS', 2670),
 ('GREEN', 2622),
 ('WORLDPOST', 2579),
 ('RELIGION', 2556),
 ('STYLE', 2254),
 ('SCIENCE', 2178),
 ('WORLD NEWS', 2177),
 ('TASTE', 2096),
 ('TECH', 2082),
 ('MONEY', 1707),
 ('ARTS', 1509),
 ('FIFTY', 1401),
 ('GOOD NEWS', 1398),
 ('ARTS & CULTURE', 1339),
 ('ENVIRONMENT', 1323),
```

```
 ('COLLEGE', 1144),
 ('LATINO VOICES', 1129),
 ('CULTURE & ARTS', 1030),
 ('EDUCATION', 1004)]
```

Já podemos entender um pouco da complexidade do desafio. Além de ser um problema com 41 diferentes rótulos, ainda existe um grande desbalanceamento entre eles, sendo que a classe com mais amostras possui 32738 amostras, tendo uma grande disparidade entre várias outras classes, sendo que a que tem menos possui apenas 1004.

1 Clusterização

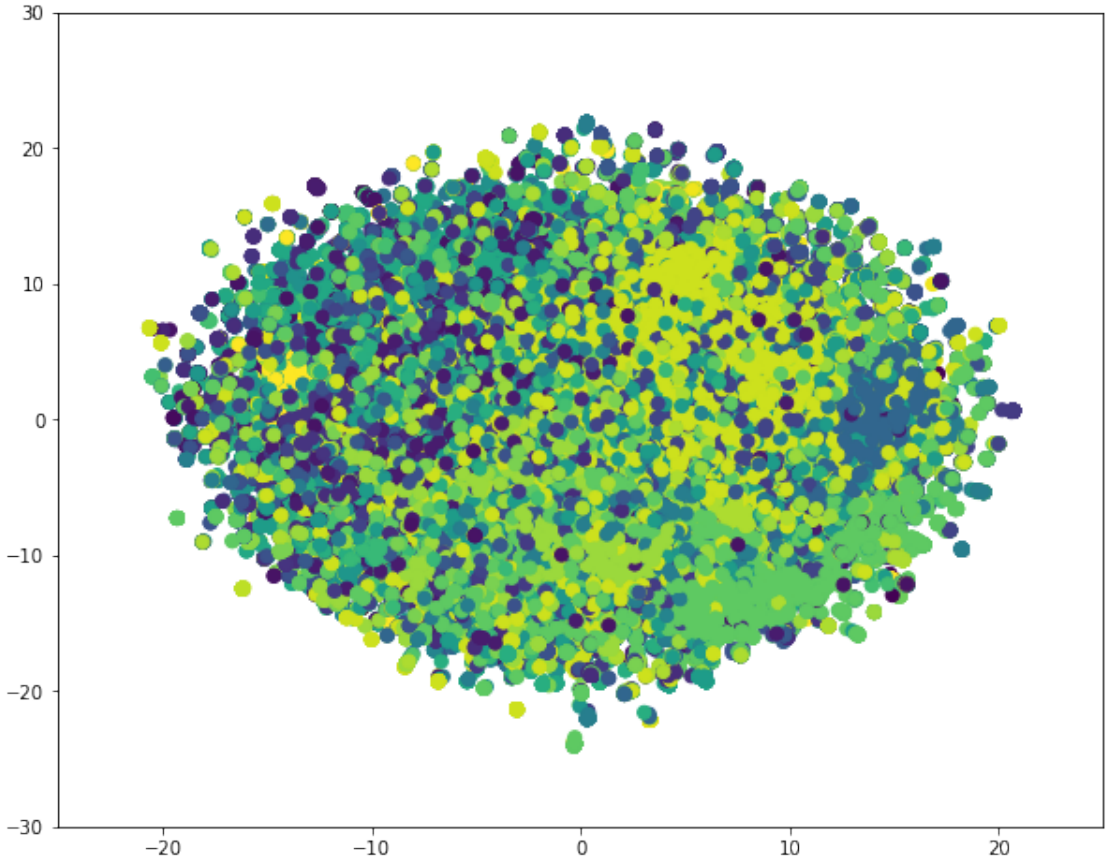
Para entender um pouco mais sobre a complexidade do problema, é interessante tentar separar o dataset em clusters e ver como as amostras estão distintas ou não entre si. Para isso, utilizaremos o TfIdf para converter os textos é dados estruturados e utilizar o t-SNE para extrair os principais componentes.

[8]:
```
def clustering(column):
    vectorizer = TfidfVectorizer()

    tfidf = vectorizer.fit_transform(df[column].values)

    embbed = TSNE(n_components=2).fit_transform(tfidf)

    encoder = LabelEncoder()

    encoded_labels = encoder.fit_transform(df['category'].values)

    plt.figure(figsize = (10,8))
    plt.scatter(embbed[:, 0], embbed[:, 1],c=encoded_labels ,s=50,␣
↪cmap='viridis')

    if column == 'headline':
        plt.ylim(-30, 30)
        plt.xlim(-25, 25)
```

1.1 Título

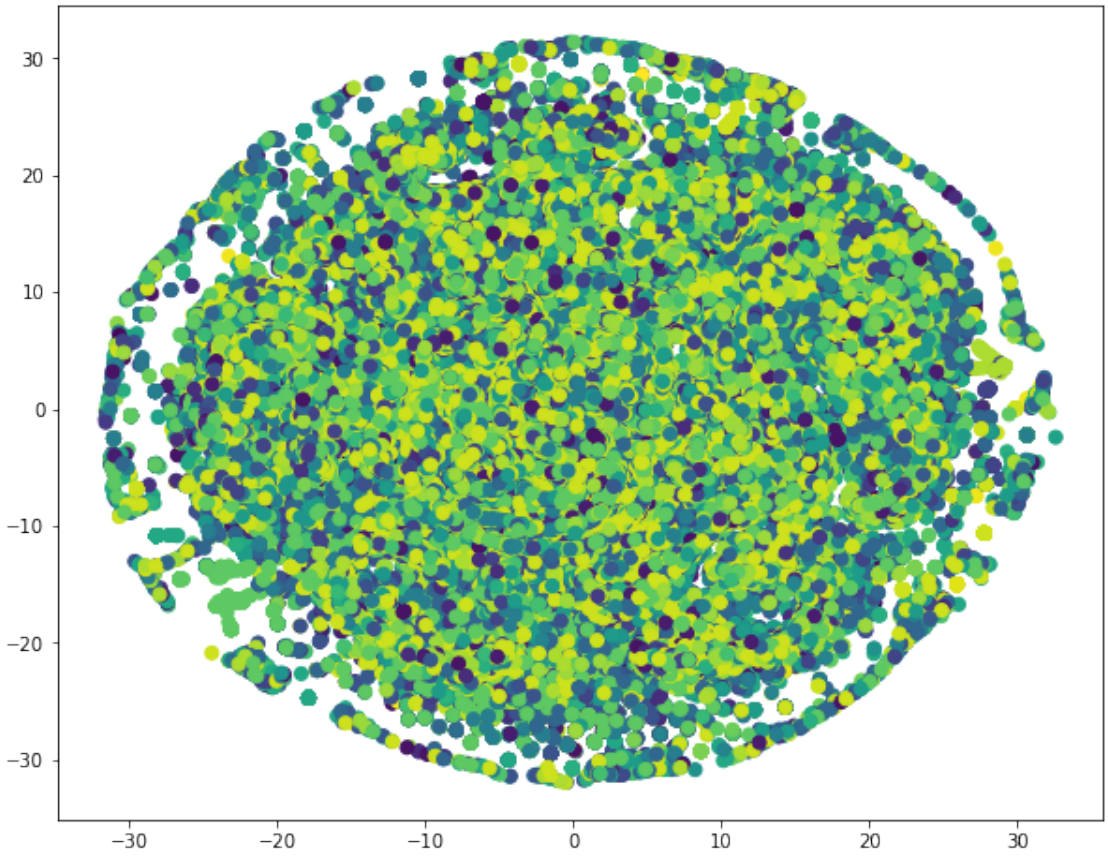
[58]:
```
clustering('headline')
```

[58]:
```
(-25, 25)
```



1.2 Descrição

[9]:
```
clustering('short_description')
```



Com a clusterização das classes, conseguimos ver quanto difícil será um modelo ter uma alta performance para este problema, uma vez que não existe uma separação bem definida entre as classes. Podemos notar que os clusters se misturam muito entre si, fazer uma separação perfeita é impossível neste caso.

2 Analisando estatísticas dos textos

[7]:
```
def column_analysis(column=None):

    print('Mínimo:',df[column].min())
    print('Máximo:',df[column].max())
    print('Média:',df[column].mean())

    plt.figure(figsize=(9,6))
    ax = df[column].plot(bins=50, kind='hist')

    df.hist(column=column, by='category', bins=50,figsize=(20,15))
```

2.1 Título

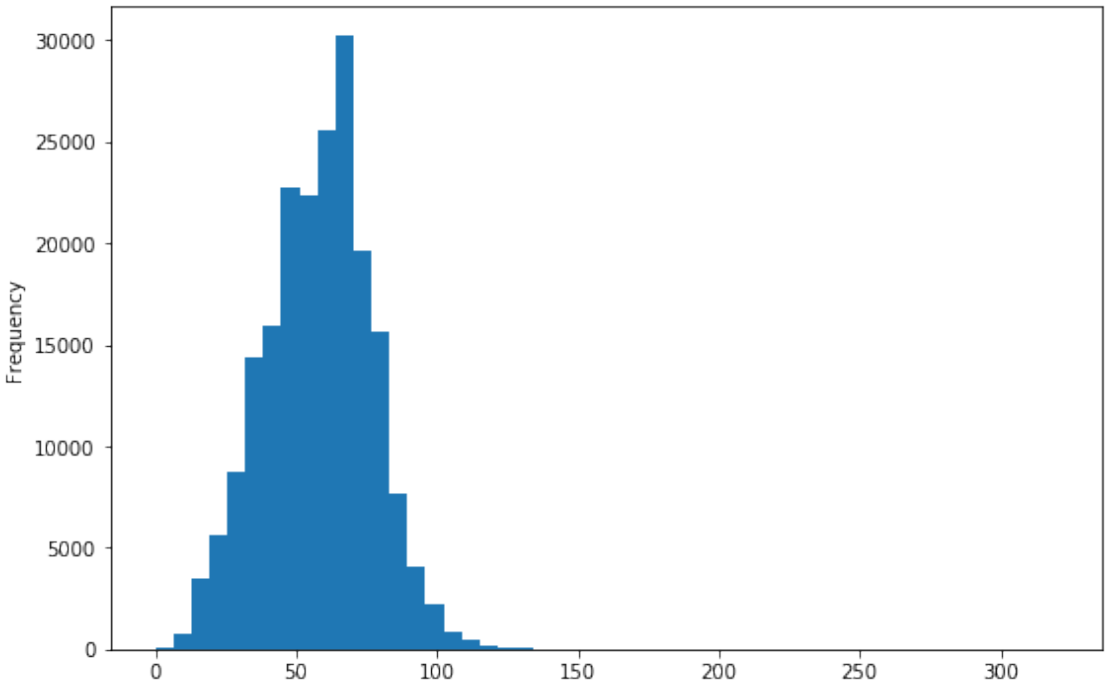
2.1.1 Nível de caracter

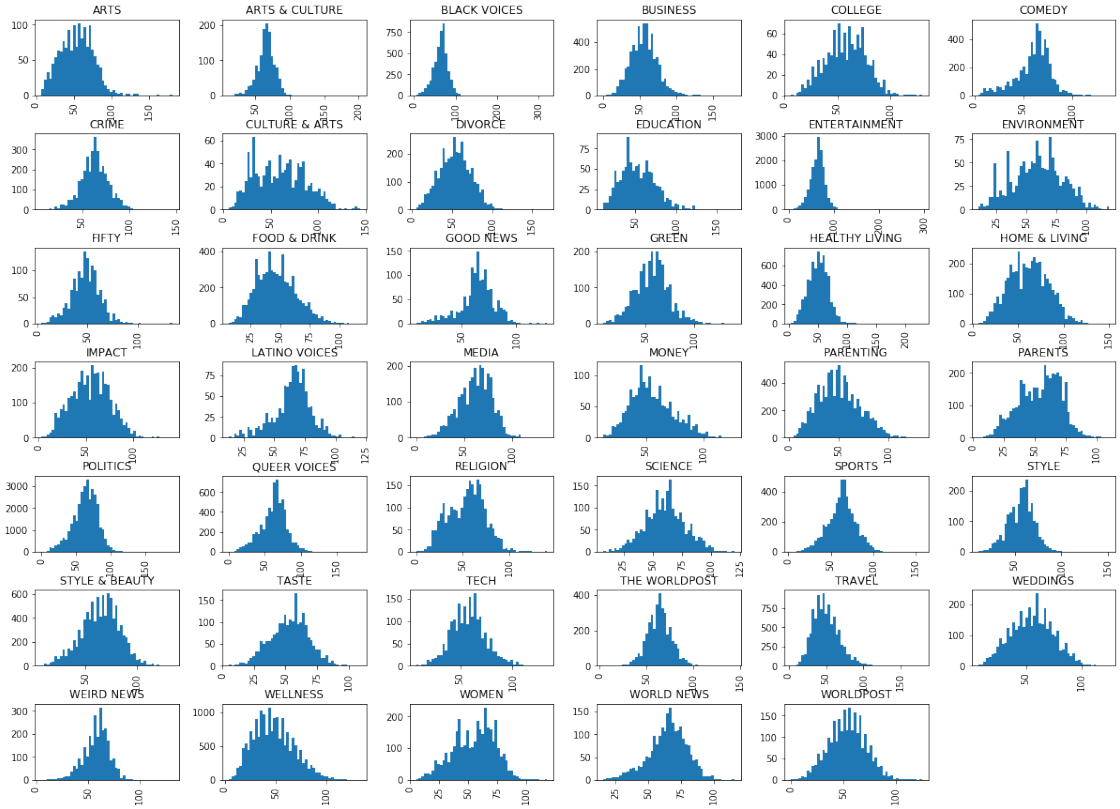
[8]:
```
df['length_headline_character'] = df['headline'].apply(len)
df.head()
```

[8]:
```
         category                                           headline  \
0          CRIME  There Were 2 Mass Shootings In Texas Last Week…
1  ENTERTAINMENT  Will Smith Joins Diplo And Nicky Jam For The 2…
2  ENTERTAINMENT    Hugh Grant Marries For The First Time At Age 57
3  ENTERTAINMENT  Jim Carrey Blasts 'Castrato' Adam Schiff And D…
4  ENTERTAINMENT  Julianna Margulies Uses Donald Trump Poop Bags…

            authors                                               link  \
0  Melissa Jeltsen  https://www.huffingtonpost.com/entry/texas-ama…
1    Andy McDonald  https://www.huffingtonpost.com/entry/will-smit…
2       Ron Dicker  https://www.huffingtonpost.com/entry/hugh-gran…
3       Ron Dicker  https://www.huffingtonpost.com/entry/jim-carre…
4       Ron Dicker  https://www.huffingtonpost.com/entry/julianna-…

                                 short_description       date  \
0  She left her husband. He killed their children… 2018-05-26
1                           Of course it has a song. 2018-05-26
2  The actor and his longtime girlfriend Anna Ebe… 2018-05-26
3  The actor gives Dems an ass-kicking for not fi… 2018-05-26
4  The "Dietland" actress said using the bags is … 2018-05-26

   length_headline_character
0                         64
1                         75
2                         47
3                         69
4                         71
```

[9]:
```
column_analysis('length_headline_character')
```

```
Mínimo: 0
Máximo: 320
Média: 57.94030460087726
```


Frequency
30000
25000
20000
15000
10000
5000
0
0
50
100
150
200
250
300


ARTS
ARTS & CULTURE
BLACK VOICES
BUSINESS
COLLEGE
COMEDY
CRIME
CULTURE & ARTS
DIVORCE
EDUCATION
ENTERTAINMENT
ENVIRONMENT
FIFTY
FOOD & DRINK
GOOD NEWS
GREEN
HEALTHY LIVING
HOME & LIVING
IMPACT
LATINO VOICES
MEDIA
MONEY
PARENTING
PARENTS
POLITICS
QUEER VOICES
RELIGION
SCIENCE
SPORTS
STYLE
STYLE & BEAUTY
TASTE
TECH
THE WORLDPOST
TRAVEL
WEDDINGS
WEIRD NEWS
WELLNESS
WOMEN
WORLD NEWS
WORLDPOST

O histograma mostra que o range do tamanho dos títulos a nível de caracter varia entre 0 e 320, tendo uma média de aproximadamente 58 caracteres. Estranho o fato de existir um título com tamanho 0, ou seja, alguma notícia não teve um título vinculado a ela.

Além disso, entre as diferentes categorias, não há alguma que se destaca por possuir um tamanho do título muito diferente das outras, em geral as categorias seguem uma distribuição parecida , com uma média realmente entre 50 e 50 caracteres

2.1.2 Nível de palavra

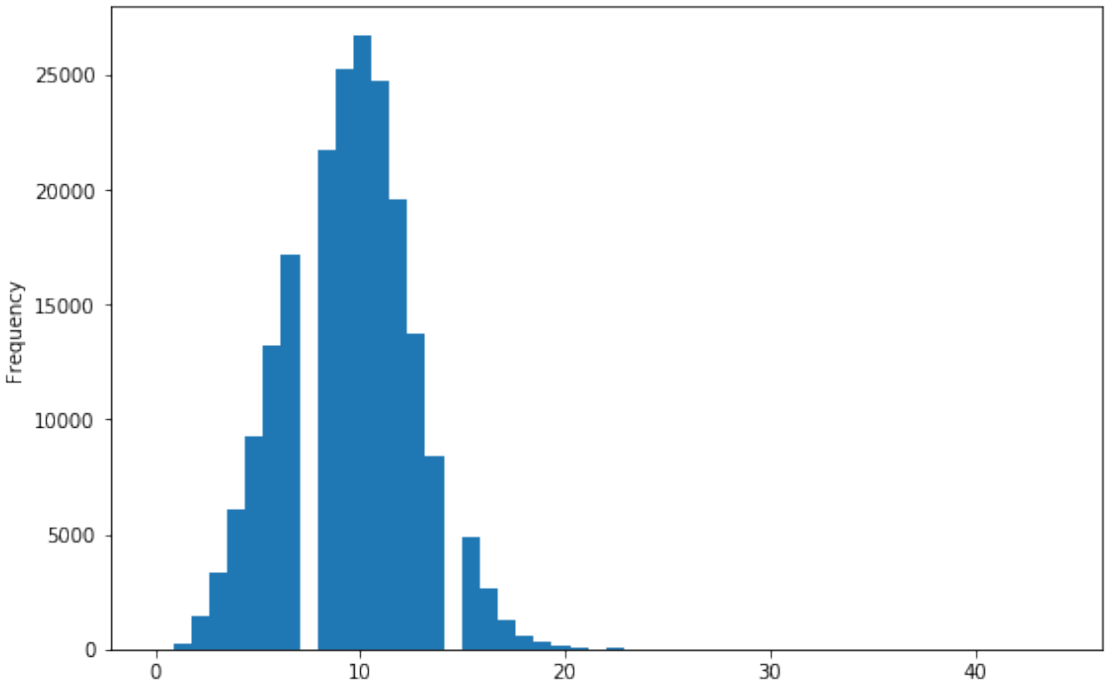
[54]:
```
df['length_headline_word'] = df['headline'].apply(lambda x: len(x.split()))
df.head()
```

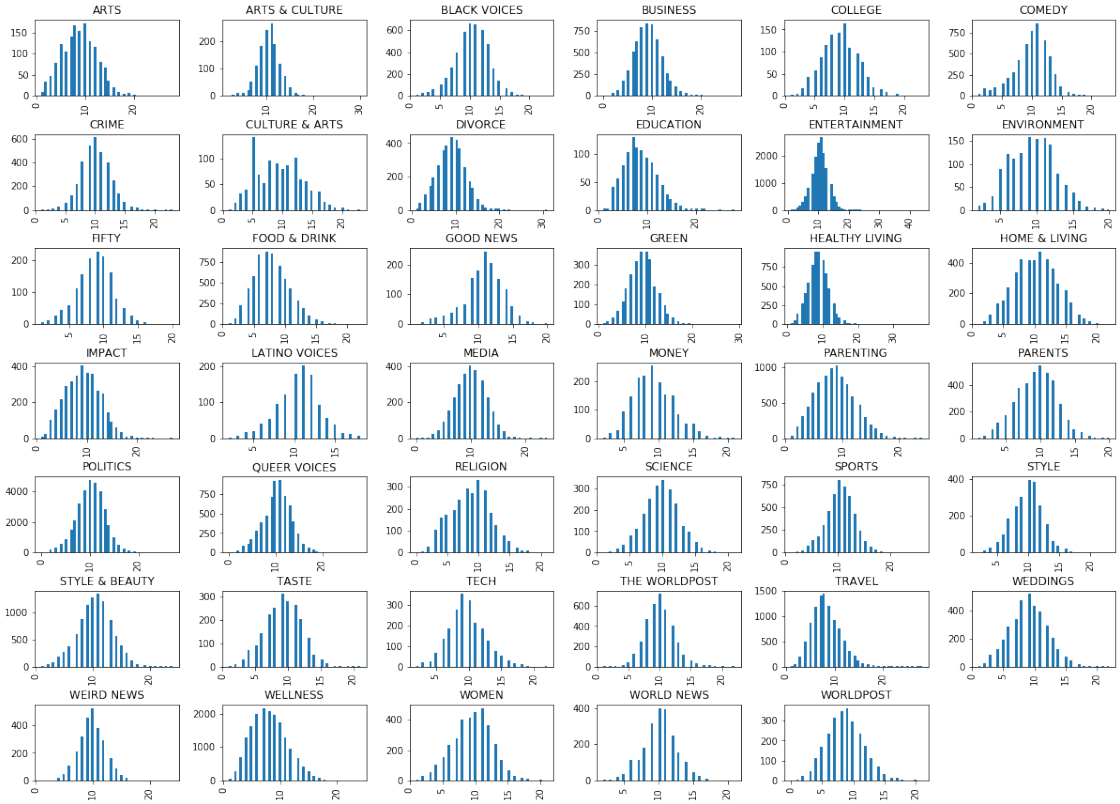
[54]:

	category	headline	authors	link	short_description	date	length_headline_character	length_headline_word
0	CRIME	There Were 2 Mass Shootings In Texas Last Week…	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-ama…	She left her husband. He killed their children…	2018-05-26	64	14
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2…	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit…	Of course it has a song.	2018-05-26	75	14
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran…	The actor and his longtime girlfriend Anna Ebe…	2018-05-26	47	10
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D…	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carre…	The actor gives Dems an ass-kicking for not fi…	2018-05-26	69	11
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags…	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-…	The "Dietland" actress said using the bags is …	2018-05-26	71	13

[55]:
```
column_analysis('length_headline_word')
```

```
Mínimo: 0
Máximo: 44
Média: 9.538563028682669
```


Frequency
25000
20000
15000
10000
5000
0
0
10
20
30
40


ARTS
ARTS & CULTURE
BLACK VOICES
BUSINESS
COLLEGE
COMEDY
CRIME
CULTURE & ARTS
DIVORCE
EDUCATION
ENTERTAINMENT
ENVIRONMENT
FIFTY
FOOD & DRINK
GOOD NEWS
GREEN
HEALTHY LIVING
HOME & LIVING
IMPACT
LATINO VOICES
MEDIA
MONEY
PARENTING
PARENTS
POLITICS
QUEER VOICES
RELIGION
SCIENCE
SPORTS
STYLE
STYLE & BEAUTY
TASTE
TECH
THE WORLDPOST
TRAVEL
WEDDINGS
WEIRD NEWS
WELLNESS
WOMEN
WORLD NEWS
WORLDPOST

O histograma mostra que o range do tamanho dos títulos a nível de palavra varia entre 0 e 44, tendo uma média de aproxidamente 9 palavras.

Em relação a distribuição entre as categorias, podemos notar que a categoria entertainment possui um padrão mais definido quanto ao número de palavras por título, é a única classe, junto com healthy living talvez, que possui uma distribuição mais densa, concentrando a maioria dos títulos com tamanho entre 10 e 15 palavras.

2.1.3 Analisando Stopwords

[9]:
```
stopwords = nltk.corpus.stopwords.words('english')
```

[31]:
```
print(stopwords[:10])
```

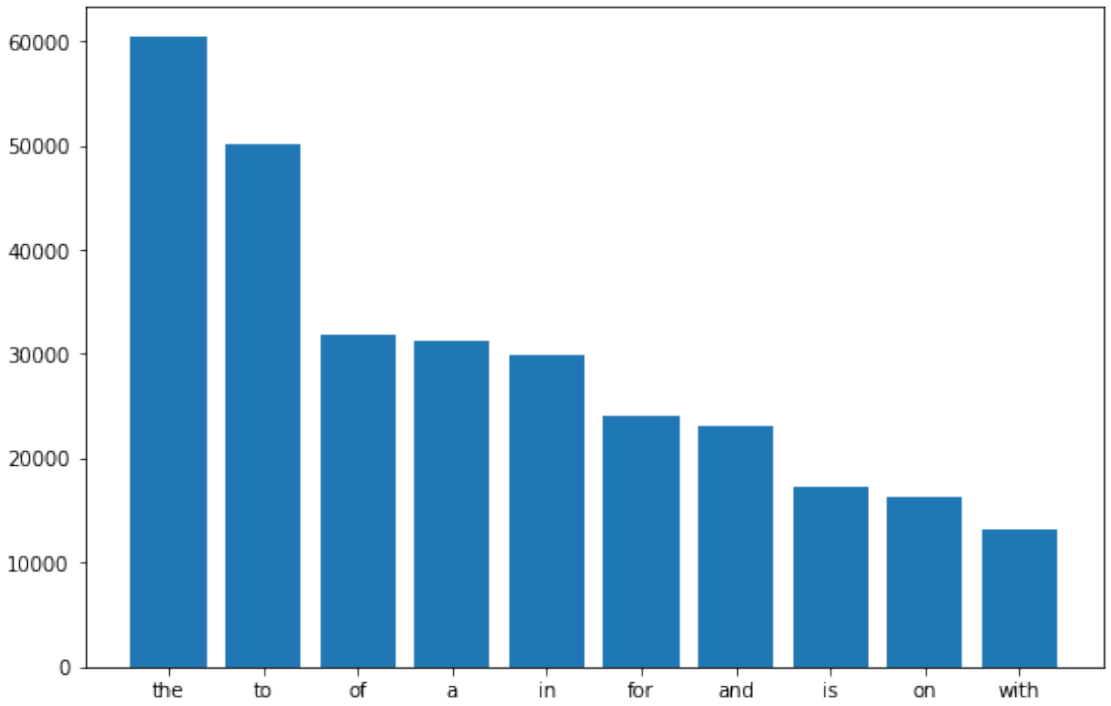
```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"]
```

Geralmente em títulos de notícias costuma-se usar muitas stopwords, que são palavras "comuns" como apresentado acima. Iremos ver quais são as mais frequentes em nosso corpus.

[69]:
```
corpus=[]
new = df['headline'].str.split()
new = new.values.tolist()
corpus = [word.lower() for i in new for word in i]

dict_stop = {}

for word in corpus:
    if word in stopwords:
        if word not in dict_stop:
            dict_stop[word] = 0
        dict_stop[word] += 1
```

[74]:
```
top = sorted(dict_stop.items(), key=lambda x:x[1],reverse=True)[:10]
x, y =z ip(*top)
plt.figure(figsize=(9,6))
plt.bar(x,y)
```

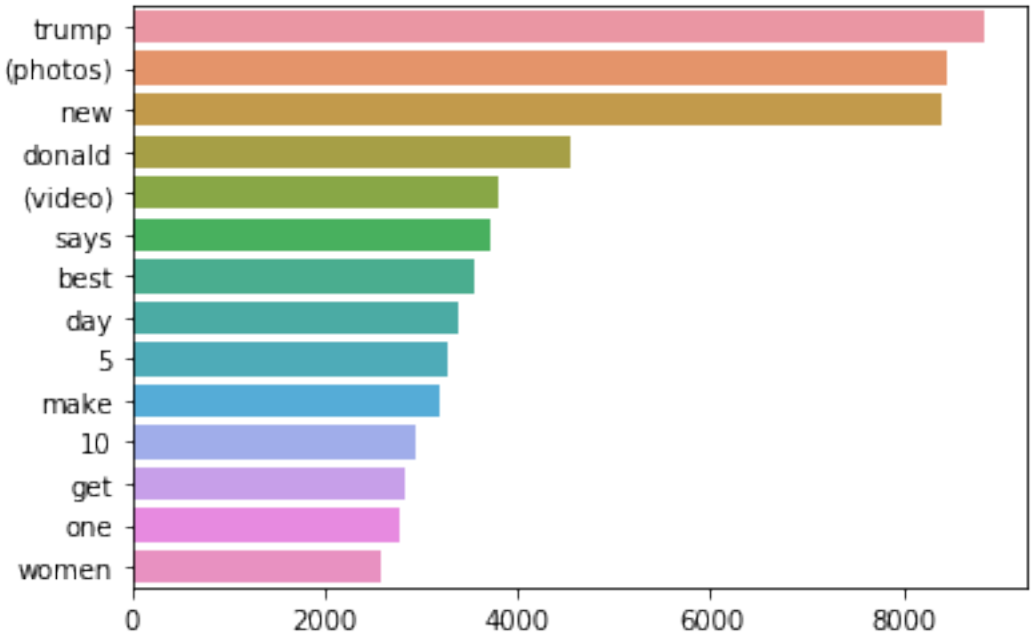
[74]:
```
<BarContainer object of 10 artists>
```



É possível notar que as stopwords “the”, “to”, “of”, “a” e “in” predominam nos títulos das notícias. Agora que sabemos que existem muitas stopwords presentes nos títulos, vamos verificar outras palavras que se estão muito presentes.

[79]:
```
most_common = Counter(corpus).most_common()

x, y= [], []

for word,count in most_common[:60]:
    if (word not in stopwords):
        x.append(word)
        y.append(count)

sns.barplot(x=y,y=x)
```

[79]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7f3c5e5f10>
```



Aqui podemos ver que Donald Trump é predominante nas notícias do nosso dataset, o que comprova que a maioria das amostras realmente são de política. New acredito que seja uma palavra irrelevante para ajudar o classificador a diferencias as categorias, assim como (photos) ou (videos). Vale o teste quando formos fazer o pré-processamento.

2.1.4 Por categoria

[26]:
```
fig, axs = plt.subplots(11, 4, figsize=(20,60))
plt.subplots_adjust(wspace=0.3, hspace=0.3)
line = 0
coluna = 0
for category in df['category'].unique():

    corpus=[]
    new = df[df['category'] == category]['headline'].str.split()
    new = new.values.tolist()
    corpus = [word.lower() for i in new for word in i]

    most_common = Counter(corpus).most_common()

    x, y= [], []

    for word,count in most_common[:40]:
        if (word not in stopwords):
            x.append(word)
            y.append(count)
```

```
    sns.barplot(x=y,y=x, ax=axs[line, coluna])

    try:
        axs[line, coluna].set_ylabel('')
        axs[line, coluna].set_xlabel('')
        axs[line, coluna].set_title(category)
    except:
        pass

    coluna+=1
    if coluna > 3:
        coluna=0
        line+=1
```


CRIME
police
man
shooting
cops
video
allegedly
accused
woman
suspect
arrested
found
death
new
shot
dead
murder
sex
boy
cop
7
officer
killing
charged
ENTERTAINMENT
new
the
trump
says
first
donald
movie
trailer
&
WORLD NEWS
north
korea
u.s.
trump
says
new
attack
south
police
president
un
killed
world
rohingya
london
dead
nuclear
east
crisis
may
IMPACT
homeless
world
help
day
new
women
people
health
one
POLITICS
trump
donald
trump's
gop
clinton
says
hillary
new
obama
house
u.s.
white
court
republicans
WEIRD NEWS
men
woman
police
dog
watch
new
gets
cops
people
weird
cat
video
BLACK VOICES
black
new
police
white
says
[photos]
women
man
WOMEN
women
woman
tweets
sexual
week
funniest
women's
men
trump
20
COMEDY
trump
donald
[video]
colbert
jimmy
trump's
stephen
fallon
snl
new
maher
seth
trevor
the
john
QUEER VOICES
gay
new
queer
lgbt
trans
transgender
lgbtq
marriage
people
here's
men
drag
SPORTS
nfl
game
[video]
world
olympic
football
nba
player
team
u.s.
bowl
super
new
win
first
olympics
says
sports
BUSINESS
business
women
wall
new
24/7
st
10
one
u.s.
q&a
TRAVEL
[photos]
travel
best
10
new
5
world
hotel
[video]
hotels
world's
top
vacation
places
around
7
things
MEDIA
trump
news
new
donald
fox
media
says
york
times
cnn
bill
trump's
journalists
reporter
post
debate
TECH
apple
new
facebook
google
iphone
week
watch
tech
twitter
top
youtube
9
videos
could
internet
RELIGION
pope
daily
meditation
francis
church
muslim
new
religious
faith
christian
catholic
religion
world
SCIENCE
new
may
scientists
space
study
nasa
science
[video]
shows
earth
could
mars
say
here's
suggests
LATINO VOICES
latino
latinos
trump
puerto
new
says
latina
u.s.
video
donald
immigration
EDUCATION
school
education
schools
teachers
students
teacher
global
student
new
college
search
education
learning
public
charter
teaching
COLLEGE
college
students
university
student
sexual
campus
colleges
assault
rape
new
school
education
PARENTS
kids
mom
parents
baby
parenting
moms
dad
day
new
daughter
things
ARTS & CULTURE
art
new
book
artist
trump
women
artists
photos
books
STYLE
fashion
new
beauty
best
style
dress
hair
week
looks
look
red
like
makeup
GREEN
climate
change
dog
california
new
u.s.
could
oil
TASTE
recipes
make
food
best
new
10
5
cooking
day
summer
get
easy
HEALTHY LIVING
health
5
life
ways
new
sleep
people
could
10
THE WORLDPOST
u.s.
says
attack
trump
isis
north
syria
new
korea
syrian
china
south
president
russia
killed
turkey
world
police
kills
people
refugees
election
GOOD NEWS
dog
men
day
help
homeless
life
gets
get
woman
new
day
kids
family
dogs
WORLDPOST
isis
u.s.
world
new
war
iran
china
israel
photos
ukraine
iraq
FIFTY
5
life
things
50
retirement
ways
7
best
love
age
10
ARTS
art
artist
new
first
nighter
world
10
stage
artists
make
street
one
WELLNESS
study
health
new
5
life
could
sleep
cancer
ways
test
healthy
PARENTING
kids
[video]
baby
parents
children
mom
parenting
child
new
day
HOME & LIVING
[photos]
home
[video]
house
make
diy
clean
craft
new
5
day
ways
vintage
ideas
7
best
clean
design
tips
STYLE & BEAUTY
[photos]
fashion
style
&
week
new
[photo]
dress
[video]
beauty
look
kate
hair
patti
[photos,
best
day
us
2012
red
DIVORCE
divorce
divorce
marriage
divorced
[video]
single
5
dating
ex
family
WEDDINGS
wedding
marriage
[video]
[photos]
married
couples
day
5
10
weddings
wedding
ways
propose
couple
engagement
love
weddings
FOOD & DRINK
[photos]
recipes
best
food
recipe
make
10
[video]
diy
new
5
recipes
day
eat
[photo]
ways
cooking
chocolate
cake
MONEY
credit
tax
money
financial
ways
new
pay
5
debt
things
10
get
card
best
bank
ENVIRONMENT
photos
climate
animal
[video]
week
[photos]
change
week
new
weather
extreme
3
day
baby
green
u.s.
oil
could
hurricane
animals
CULTURE & ARTS
[photos]
art
imagenblog
week
&
new
artist
arts
[video]
art
gallery
culture
artists
first
happy
exhibition
photos
museum
stage

Fazer essa análise por categoria é importante para selecionar algumas palavras que podem ser incluídas nas stopwords para serem removidas, e as vezes algumas palavras que pensavamos serem recorrentes, na verdade pode ser bastante significativa para determinada classe, como exemplo Donald Trump, que é a palavra que mais aparece, porém apenas das notícias políticas, ou seja, remover essas palavras talvez não seja uma boa ideia.

2.1.5 Wordcloud

Nuvem de palavras é uma maneira de representar textos. O tamanho e a cor de cada palavra que aparece na nuvem indicam sua frequência ou importância.

[26]:
```
def plot_wordcloud(text):

    stopwords = set(nltk.corpus.stopwords.words('english'))

    def _preprocess_text(text):

        corpus=[]
        for news in text:
            words = [w for w in news.split() if (w.lower() not in stopwords)]

            words = [w for w in words if len(w)>2]

            corpus.append(words)
        return corpus

    corpus=_preprocess_text(text)

    wordcloud = WordCloud(
        background_color='white',
        stopwords=set(STOPWORDS),
        max_words=100,
        max_font_size=30,
        scale=3,
        random_state=1)

    wordcloud=wordcloud.generate(str(corpus))


    plt.figure(figsize=(10,10))
    plt.axis('off')
    plt.imshow(wordcloud)
    plt.show()
```

[15]:
```
plot_wordcloud(df['headline'])
```

2.2 Descrição

2.2.1 Nível de caracter

[10]:
```
df['length_description_character'] = df['short_description'].apply(len)
df.head()
```

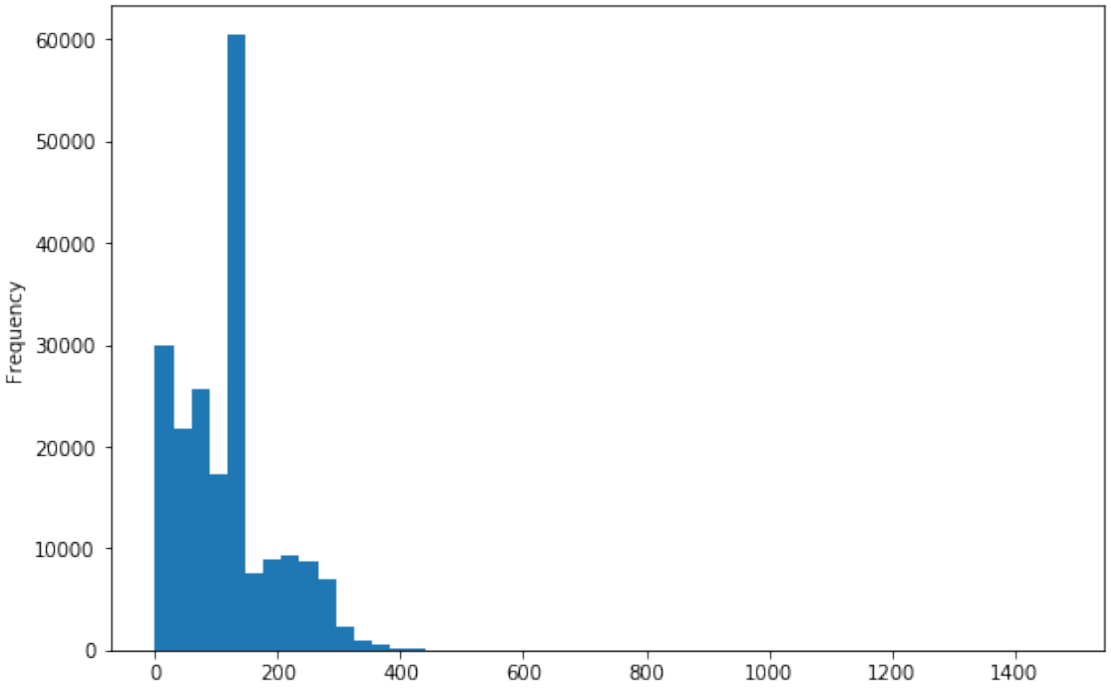
[10]:

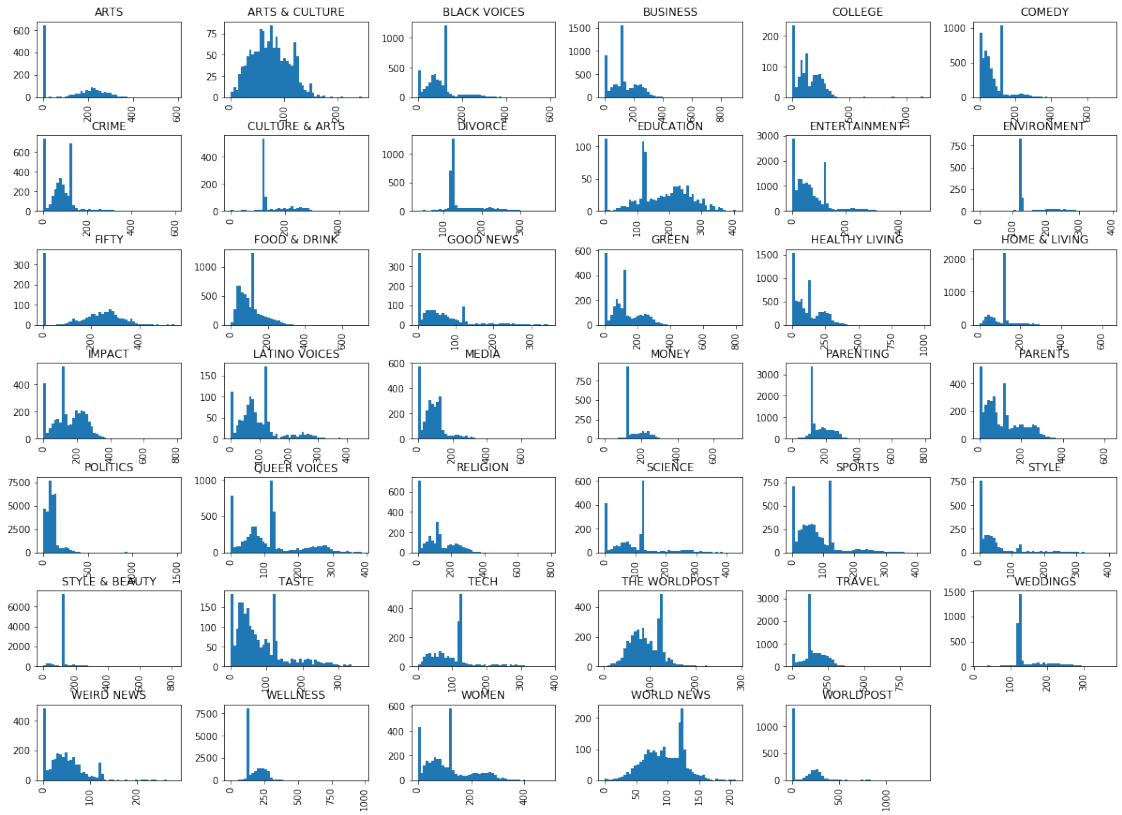
	category	headline	authors	link	short_description	date
0	CRIME	There Were 2 Mass Shootings In Texas Last Week…	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-ama…	She left her husband. He killed their children…	2018-05-26
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2…	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit…	Of course it has a song.	2018-05-26
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran…	The actor and his longtime girlfriend Anna Ebe…	2018-05-26
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D…	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carre…	The actor gives Dems an ass-kicking for not fi…	2018-05-26
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags…	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-…	The "Dietland" actress said using the bags is …	2018-05-26

	length_headline_character	length_description_character
0	64	76
1	75	24
2	47	87
3	69	86
4	71	87

[11]:
```
column_analysis('length_description_character')
```

```
Mínimo: 0
Máximo: 1472
Média: 114.31268639253584
```





Comparando o histograma do tamanho dos textos a nível de caracter da descrição curta com o título, vemos que segue uma distribuição diferente, seguindo menos um comportamento normal e possuindo uma quantidade maior de caracteres, o que era esperado, com uma média de 114 caracteres contra 57 do título.

É visível também a diferença dos histogramas das categorias entre as duas informações (título e descrição curta). Na descrição curta, existe uma frequência maior de determinados tamanhos de texto em valores específicos, ou seja, é mais padronizado do que o título.

2.2.2 Nível de palavra

[12]:
```
df['length_description_word'] = df['short_description'].apply(lambda x: len(x.
 ↪split()))
df.head()
```

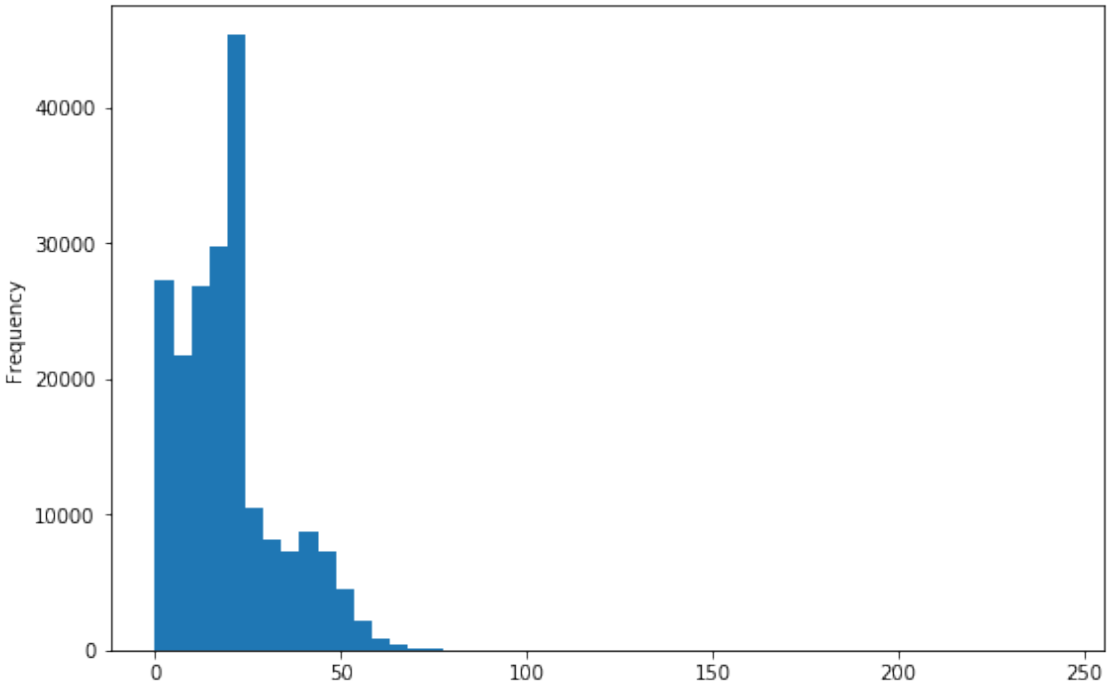
[12]:

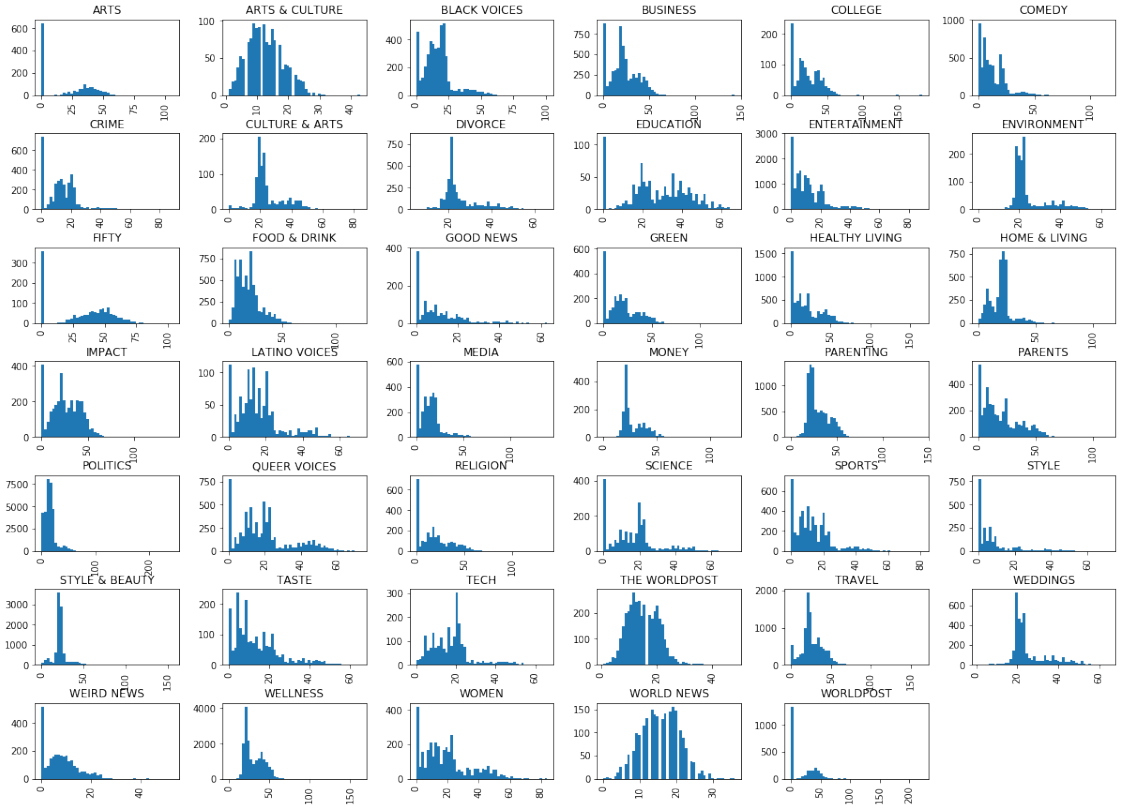
	category	headline	\
0	CRIME	There Were 2 Mass Shootings In Texas Last Week…	
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2…	
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D…	
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags…	

```
           authors                                                link  \
0  Melissa Jeltsen  https://www.huffingtonpost.com/entry/texas-ama…
1    Andy McDonald  https://www.huffingtonpost.com/entry/will-smit…
2       Ron Dicker  https://www.huffingtonpost.com/entry/hugh-gran…
3       Ron Dicker  https://www.huffingtonpost.com/entry/jim-carre…
4       Ron Dicker  https://www.huffingtonpost.com/entry/julianna-…

                                  short_description       date  \
0  She left her husband. He killed their children… 2018-05-26
1                           Of course it has a song. 2018-05-26
2  The actor and his longtime girlfriend Anna Ebe… 2018-05-26
3  The actor gives Dems an ass-kicking for not fi… 2018-05-26
4  The "Dietland" actress said using the bags is … 2018-05-26

   length_headline_character  length_description_character  \
0                         64                           76
1                         75                           24
2                         47                           87
3                         69                           86
4                         71                           87

   length_description_word
0                       13
1                        6
2                       15
3                       14
4                       13
```

[13]:
```
column_analysis('length_description_word')
```

```
Mínimo: 0
Máximo: 243
Média: 19.728288848063013
```


Frequency
40000
30000
20000
10000
0
0
50
100
150
200
250


ARTS
ARTS & CULTURE
BLACK VOICES
BUSINESS
COLLEGE
COMEDY
CRIME
CULTURE & ARTS
DIVORCE
EDUCATION
ENTERTAINMENT
ENVIRONMENT
FIFTY
FOOD & DRINK
GOOD NEWS
GREEN
HEALTHY LIVING
HOME & LIVING
IMPACT
LATINO VOICES
MEDIA
MONEY
PARENTING
PARENTS
POLITICS
QUEER VOICES
RELIGION
SCIENCE
SPORTS
STYLE
STYLE & BEAUTY
TASTE
TECH
THE WORLDPOST
TRAVEL
WEDDINGS
WEIRD NEWS
WELLNESS
WOMEN
WORLD NEWS
WORLDPOST

O histograma mostra que o range do tamanho das descrições curtas a nível de palavra varia entre 0 e 243, tendo uma média de aproximadamente 20 palavras. Seguindo a mesma idéia da análise anterior, vemos que a quantidade de palavras tem um padrão mais definido, concentrando na maioria das classes em valores menos.

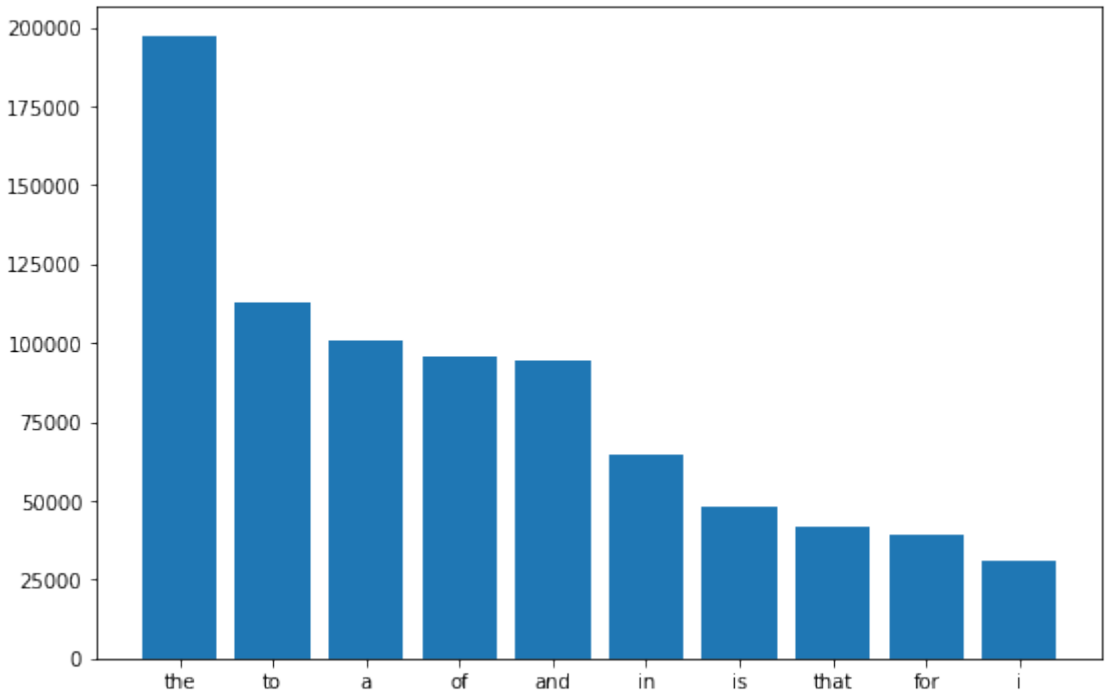
2.2.3 Analisando Stopwords

[14]:
```
stopwords = nltk.corpus.stopwords.words('english')
```

[16]:
```
corpus=[]
new = df['short_description'].str.split()
new = new.values.tolist()
corpus = [word.lower() for i in new for word in i]

dict_stop = {}

for word in corpus:
    if word in stopwords:
        if word not in dict_stop:
            dict_stop[word] = 0
        dict_stop[word] += 1
```

[19]:
```
top = sorted(dict_stop.items(), key=lambda x:x[1],reverse=True)[:10]
x, y = zip(*top)
plt.figure(figsize=(9,6))
plt.bar(x,y)
```

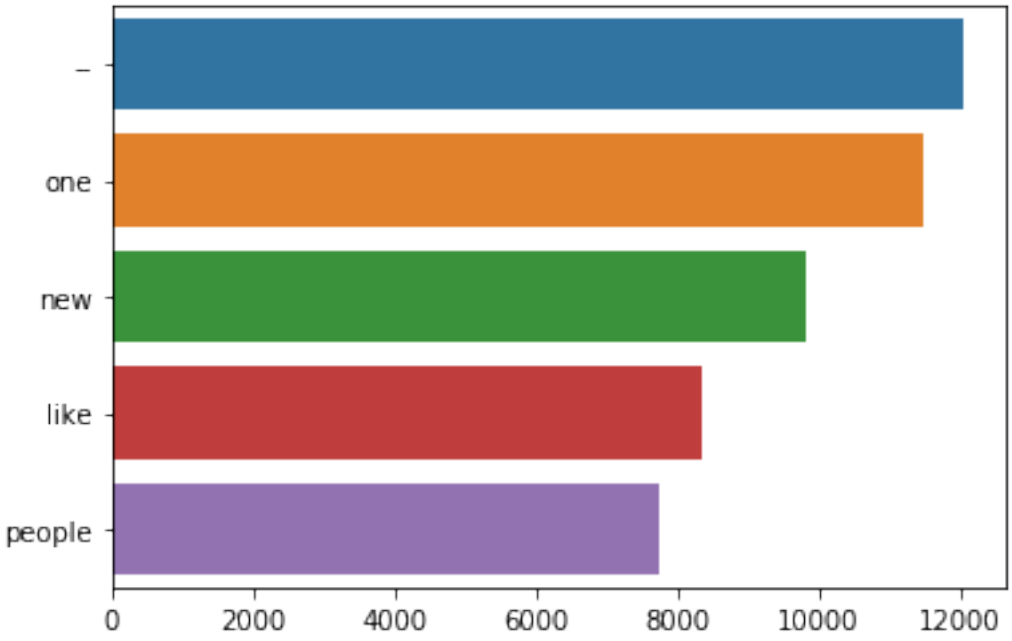
[19]:
```
<BarContainer object of 10 artists>
```



É possível notar que as stopwords “the”, “to”, “a”, “of” e “and” predominam nas descrições das notícias. Diferencia do título apenas em relação ao “in” que entrou no lugar do “and”.

[21]:
```
most_common = Counter(corpus).most_common()

x, y= [], []

for word,count in most_common[:60]:
    if (word not in stopwords):
        x.append(word)
        y.append(count)

sns.barplot(x=y,y=x)
```

[21]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1421037710>
```



Interessante notar que a palavra mais comum na descrição curta, sem contar as stopwords, é o hífen. Além disso, entre as 60 palavras mais comuns, apenas 5 não são stopwords, enquanto no título eram 14.

2.2.4 Por categoria

[24]:
```
fig, axs = plt.subplots(11, 4, figsize=(20,60))
plt.subplots_adjust(wspace=0.3, hspace=0.3)
line = 0
coluna = 0
for category in df['category'].unique():

    corpus=[]
    new = df[df['category'] == category]['short_description'].str.split()
    new = new.values.tolist()
    corpus = [word.lower() for i in new for word in i]

    most_common = Counter(corpus).most_common()

    x, y= [], []

    for word,count in most_common[:40]:
        if (word not in stopwords):
            x.append(word)
            y.append(count)
```

```
    sns.barplot(x=y,y=x, ax=axs[line, coluna])

    try:
        axs[line, coluna].set_ylabel('')
        axs[line, coluna].set_xlabel('')
        axs[line, coluna].set_title(category)
    except:
        pass

    coluna+=1
    if coluna > 3:
        coluna=0
        line+=1
```


CRIME
police
said
one
people
ENTERTAINMENT
new
one
WORLD NEWS
people
u.s.
said
president
one
IMPACT
people
one
POLITICS
trump
president
donald
said
one
WEIRD NEWS
one
like
BLACK VOICES
black
new
WOMEN
women
COMEDY
like
one
QUEER VOICES
gay
people
new
SPORTS
one
find
BUSINESS
new
TRAVEL
one
new
MEDIA
news
media
new
fox
TECH
new
apple
RELIGION
people
one
SCIENCE
new
one
may
scientists
LATINO VOICES
latino
one
new
EDUCATION
school
education
students
COLLEGE
college
students
university
PARENTS
kids
ARTS & CULTURE
new
book
artist
STYLE
like
new
GREEN
climate
one
people
new
TASTE
make
one
like
HEALTHY LIVING
THE WORLDPOST
people
said
one
u.s.
president
killed
GOOD NEWS
one
WORLDPOST
world
one
FIFTY
one
ARTS
art
one
new
WELLNESS
PARENTING
kids
one
HOME & LIVING
home
one
STYLE & BEAUTY
check
style
want
sure
fashion
money
twitter
huffpost
new
facebook
DIVORCE
divorce
one
WEDDINGS
wedding
weddings
one
FOOD & DRINK
one
like
food
make
MONEY
credit
money
financial
ENVIRONMENT
climate
week
animal
one
new
CULTURE & ARTS
art
new
one

Outra visão importante que podemos notar é o fato de que algumas palavras fazem muito sentido serem mais frequentes em determinadas classes, como alguns exemplos:

- Classe WOMEN:
  - women
- Classe Black Voices:
  - black
- Classe CRIME:
  - police
- Classe EDUCATION:
  - school
  - education
  - students

Essas são alguns dos exemplos de palavras que possam ser interessantes para compreendermos o domínio do nosso problema, e talvez limitar o vocabulário na hora de realizar o embedding dos textos, e representá-los de forma numérica.

2.2.5 Wordcloud

[27]:
```
plot_wordcloud(df['short_description'])
```

Conclusão

O problema em questão possui um grau de complexidade elevado por possuir inúmeras classes, o que faz com que fique complicado conseguir fazer com que a IA consiga fazer a distinção precisa entre todas elas. Além do mais, foi evidenciado o desbalanceamento entre as categorias, o que

causa um certo viés e faz com que a IA tenda a classificar mais amostras como sendo das classes com maior volumetria.

Por outro lado, conseguimos identificar alguns comportamentos e palavras representativas para cada classe, o que pode facilitar na hora do modelo fazer a distinção entre elas. Além do mais, utilizar dois campos, título e descrição breve, para fazer o treinamento, tende a ajudar na classificação, uma vez que vimos que as classes se diferem em palavras específicas nos dois campos.

Para agregar na produção da IA, também conseguimos identificar algumas palavras que se repetem em todas as classes, podendo ser agregadas na limpeza dos textos na hora do treinamento.