

Análise Exploratória

April 3, 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
%matplotlib inline
```

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

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

```
[4]:
```

	category	headline	authors \
count	200853	200853	200853
unique	41	199344	27993
top	POLITICS	Sunday Roundup	
freq	32739	90	36620
first	NaN	NaN	NaN
last	NaN	NaN	NaN

	link	short_description \
count	200853	200853
unique	200812	178353
top	https://www.huffingtonpost.com	http://testkitch...
freq	2	19712
first	NaN	NaN
last	NaN	NaN

	date
count	200853
unique	2309
top	2013-01-17 00:00:00
freq	100
first	2012-01-28 00:00:00
last	2018-05-26 00:00:00

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

```
[5]: 41
```

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

```
[6]: [('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 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))
```

1.1 Título

1.1.1 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
```

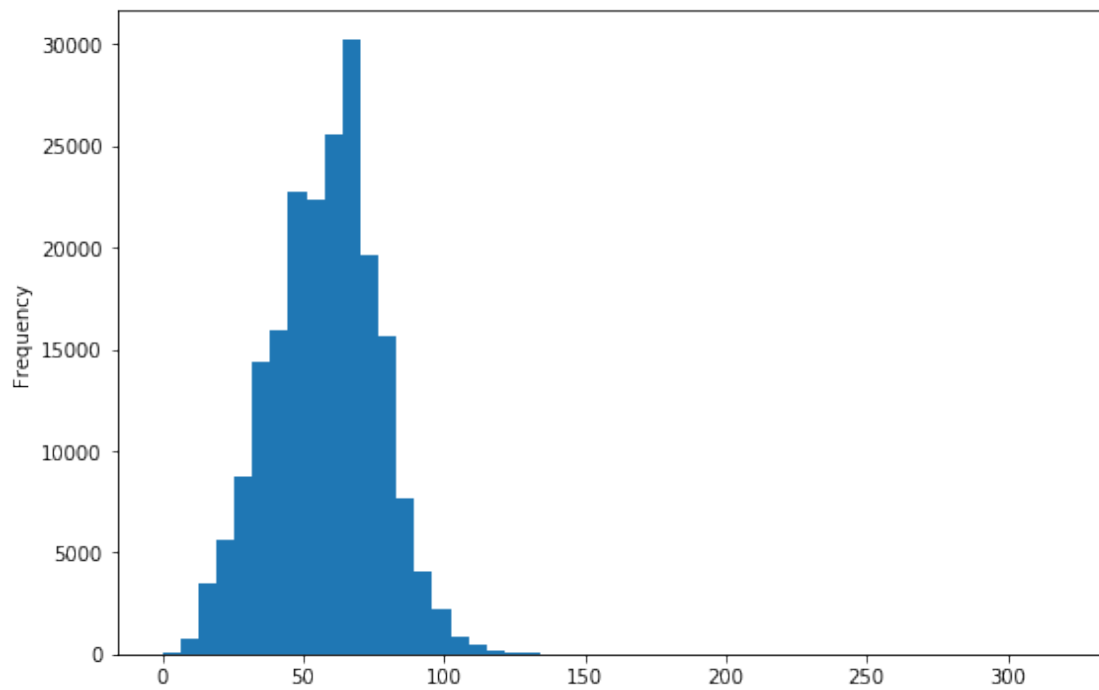
0	64
1	75
2	47
3	69
4	71

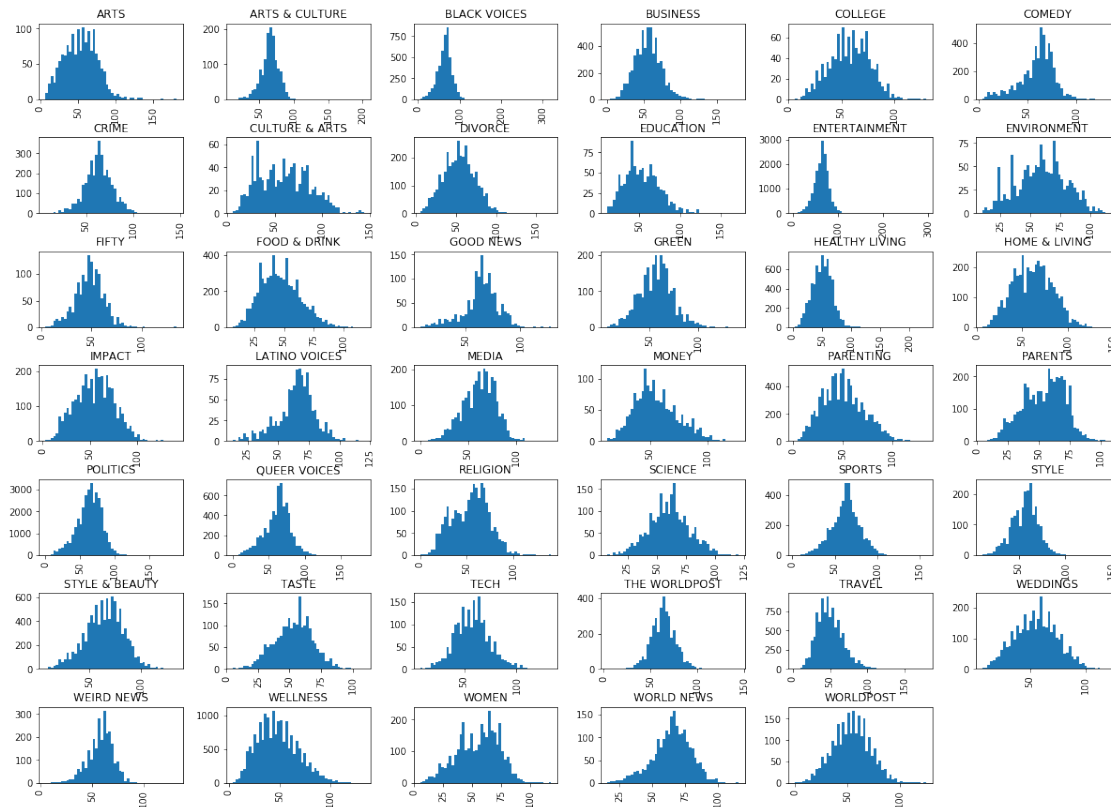
```
[9]: column_analysis('length_headline_character')
```

Mínimo: 0

Máximo: 320

Média: 57.94030460087726





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

1.1.2 Nível de palavra

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

```
[54]:
```

	category	headline \	authors	link \
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...		
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_headline_word
0                          64                   14
1                          75                   14
2                          47                   10
3                          69                   11
4                          71                   13

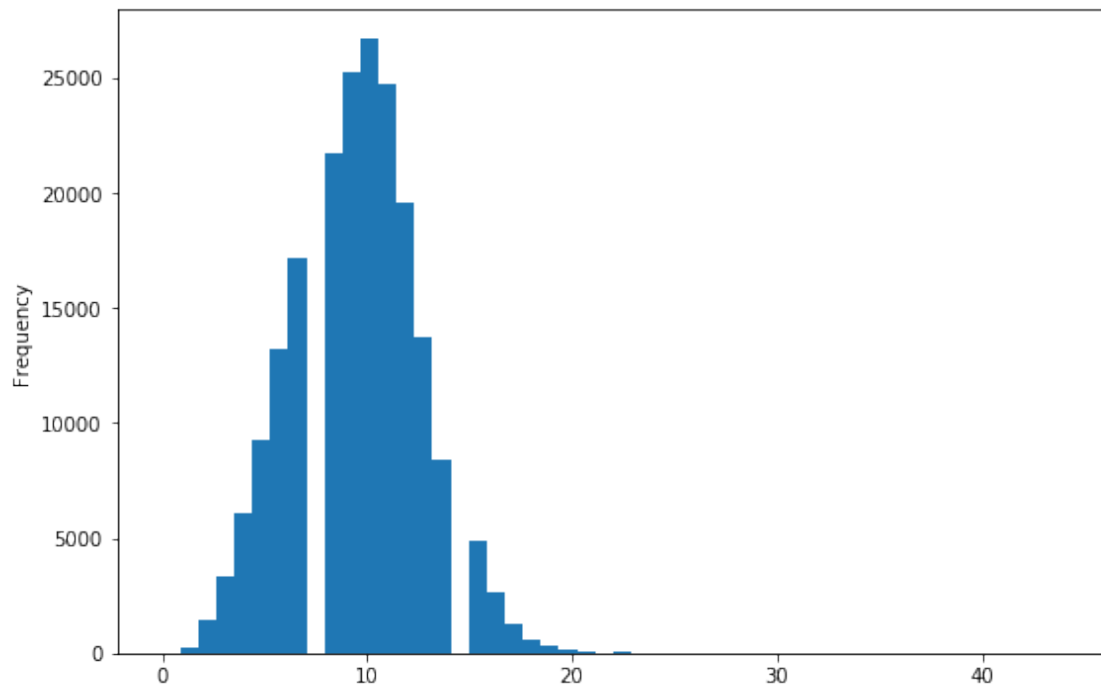
```

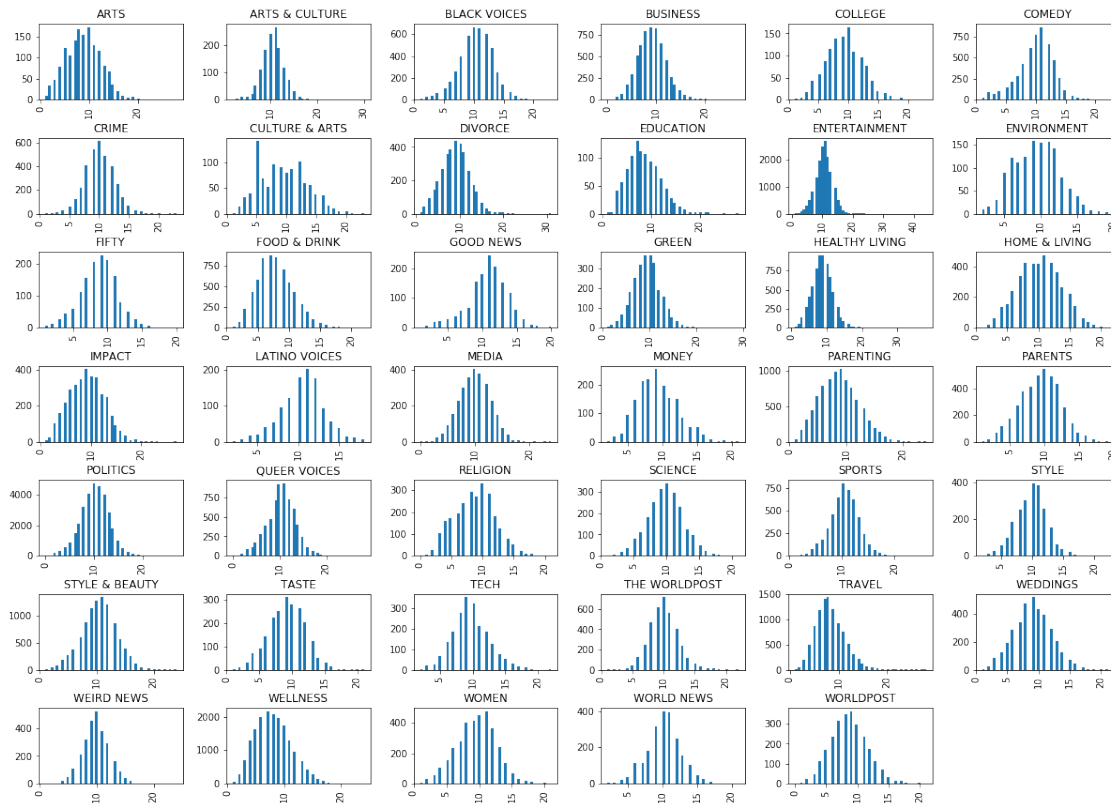
```
[55]: column_analysis('length_headline_word')
```

Mínimo: 0

Máximo: 44

Média: 9.538563028682669





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

1.1.3 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.

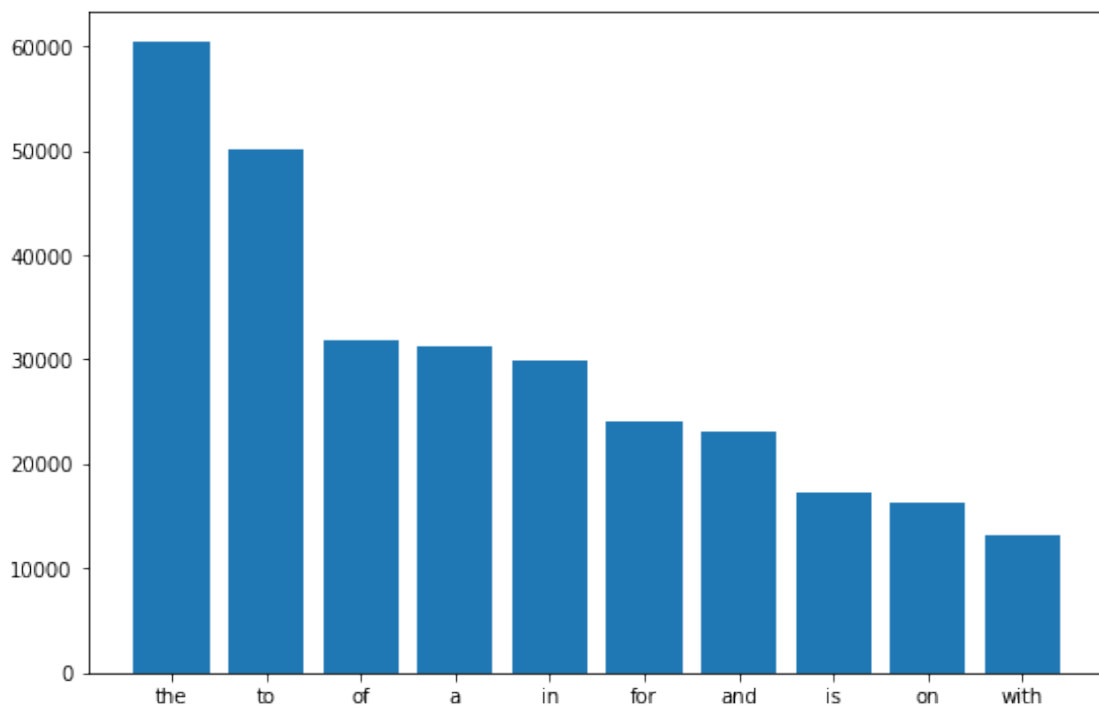
```
[69]: corpus=[]
new = df['headline'].str.split()
new = new.values.tolist()
corpus = [word.lower() for i in new for word in i]
```

```
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 = zip(*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= [], []
```



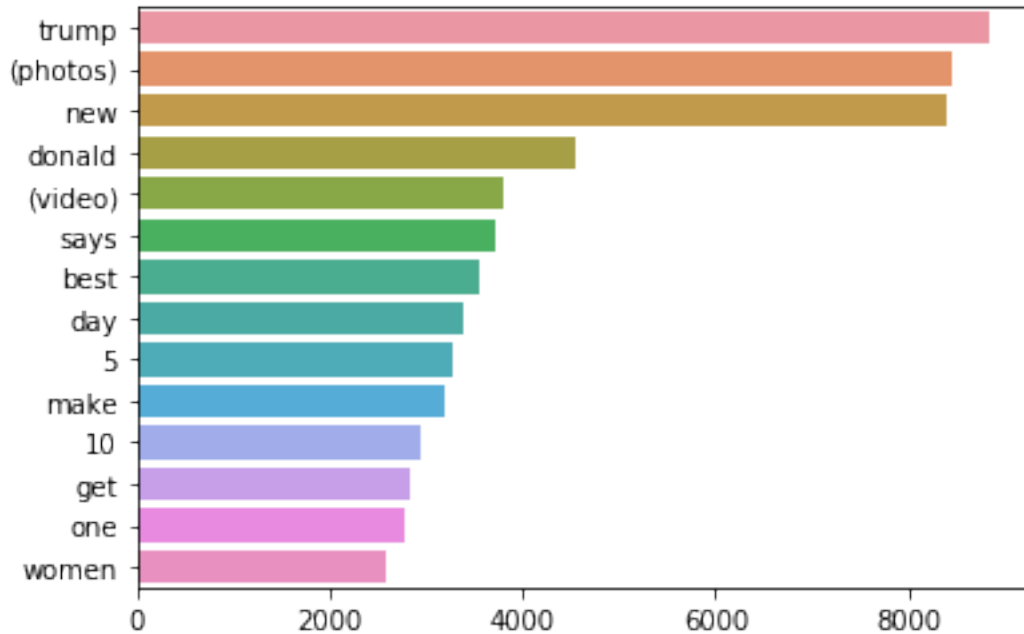
```

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 diferenciar as categorias, assim como (photos) ou (videos). Vale o teste quando formos fazer o pré-processamento.

1.1.4 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= [], []

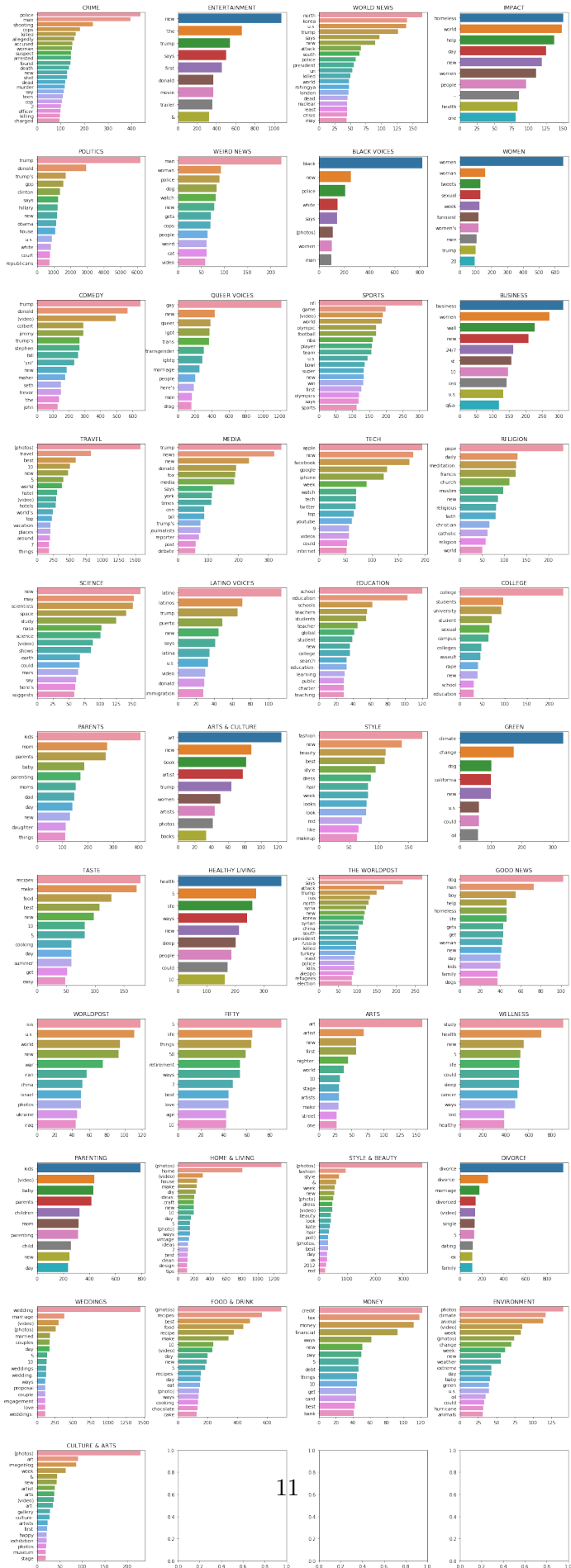
for word,count in most_common[:40]:
    if (word not in stopwords):
        x.append(word)
        y.append(count)

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

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

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

```



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 pensávamos 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.

1.1.5 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'])
```



1.2 Descrição

1.2.1 Nível de caracter

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

```
[10]:
```

	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

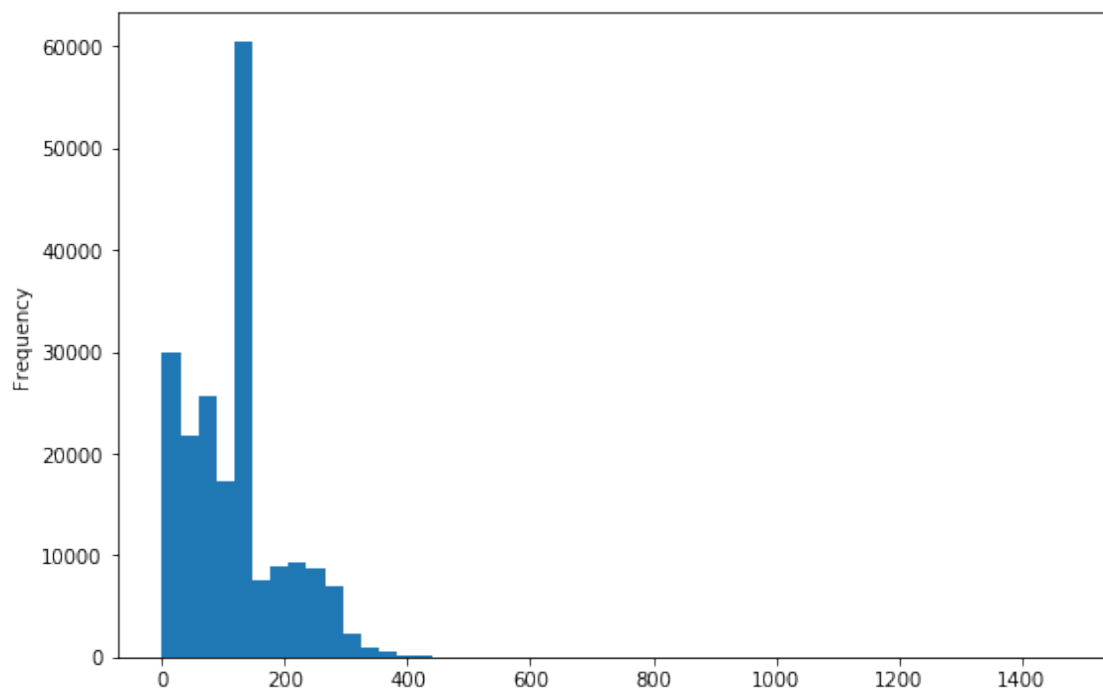
	length_headline_character	length_description_character
0	64	76
1	75	24
2	47	87
3	69	86
4	71	87

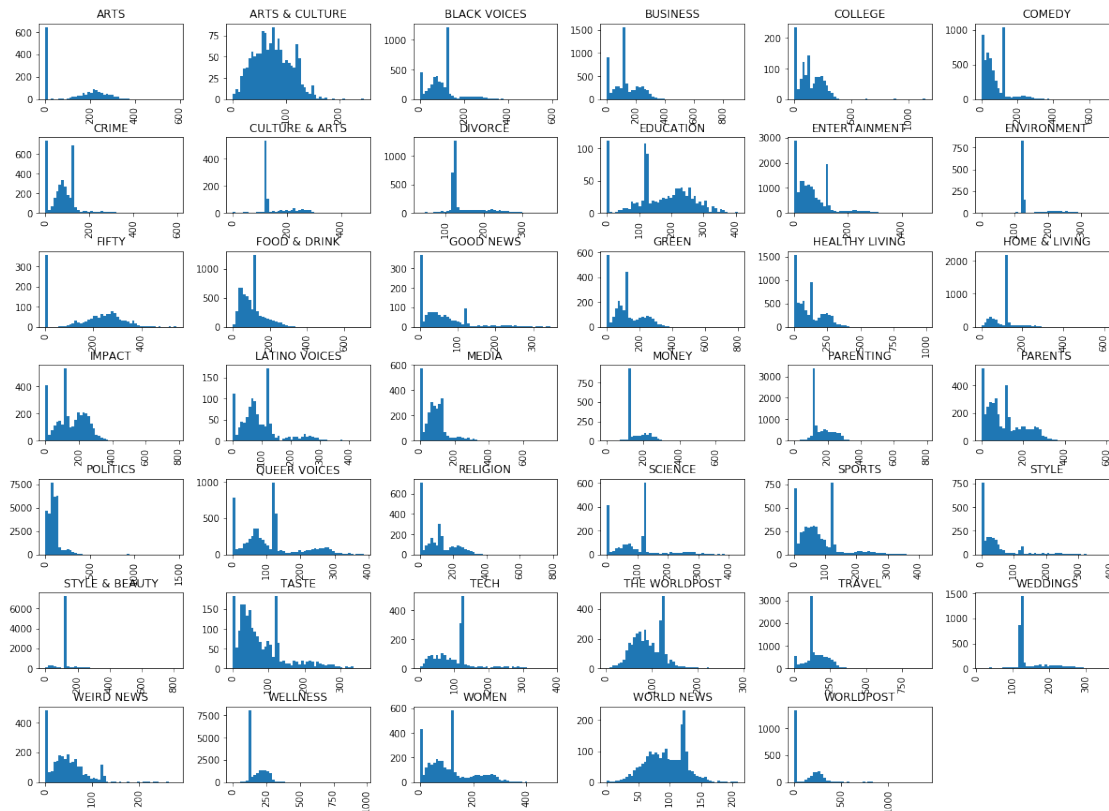
```
[11]: column_analysis('length_description_character')
```

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.

1.2.2 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-...	

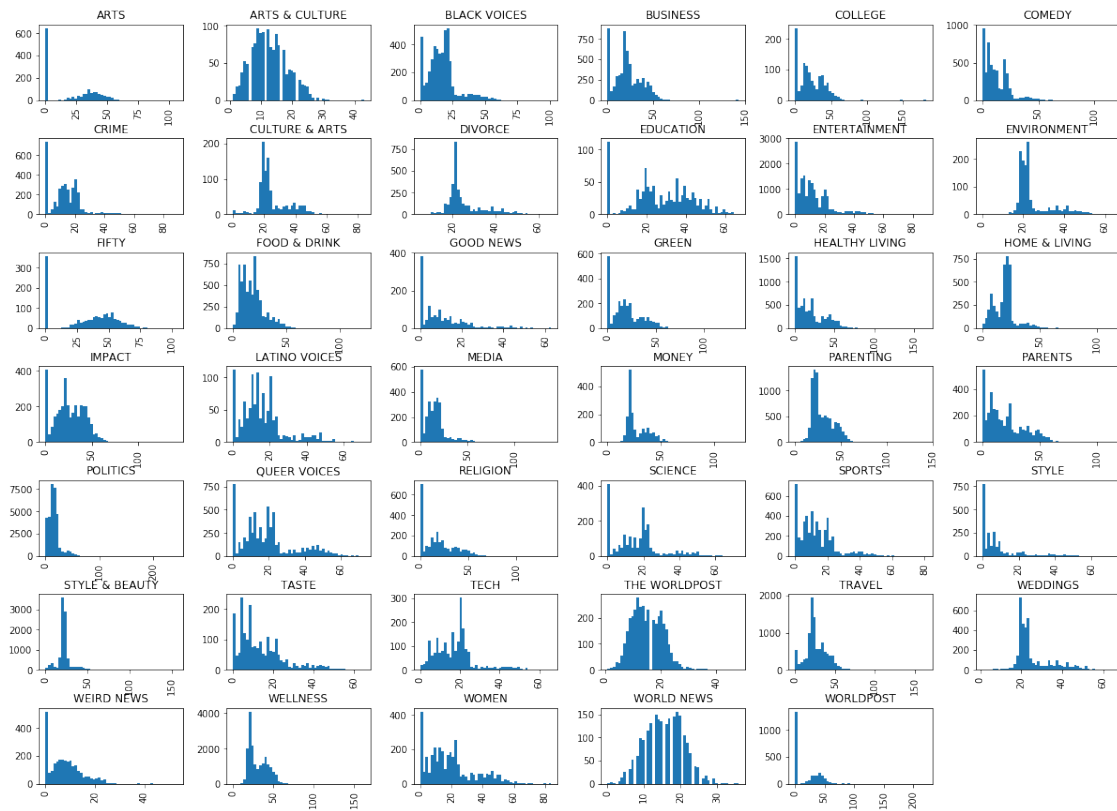
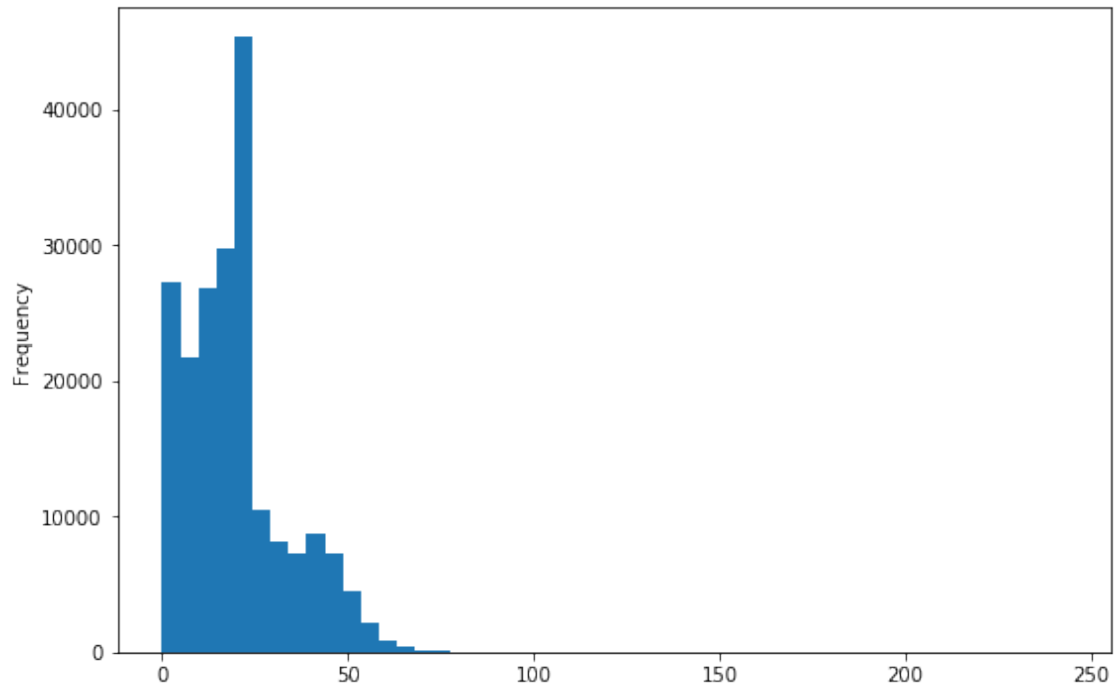
	short_description	date	\
0	She left her husband. He killed their children...	2018-05-26	
1	Of course it has a song.	2018-05-26	
2	The actor and his longtime girlfriend Anna Ebe...	2018-05-26	
3	The actor gives Dems an ass-kicking for not fi...	2018-05-26	
4	The "Dietland" actress said using the bags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

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 ideia da análise anterior, vemos que a quantidade de palavras tem um padrão mais definido, concentrando na maioria das classes em valores menos.

1.2.3 Analisando Stopwords

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

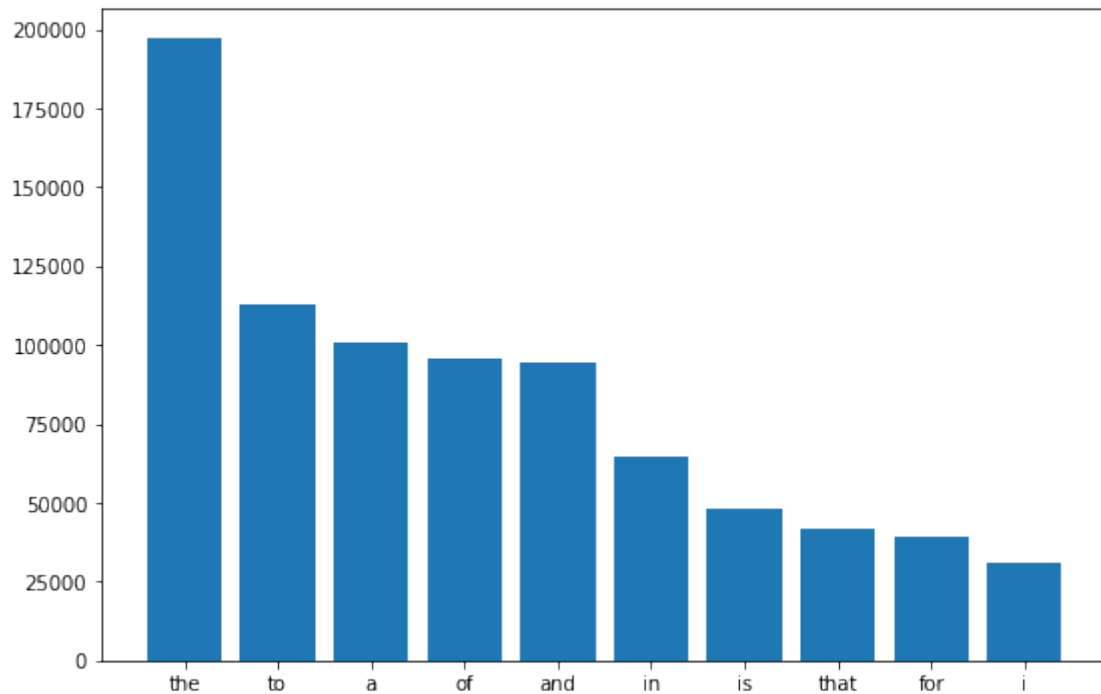
```
[16]: corpus=[]
new = df['short_description'].str.split()
new = new.values.tolist()
corpus = [word.lower() for i in new for word in i]

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”.

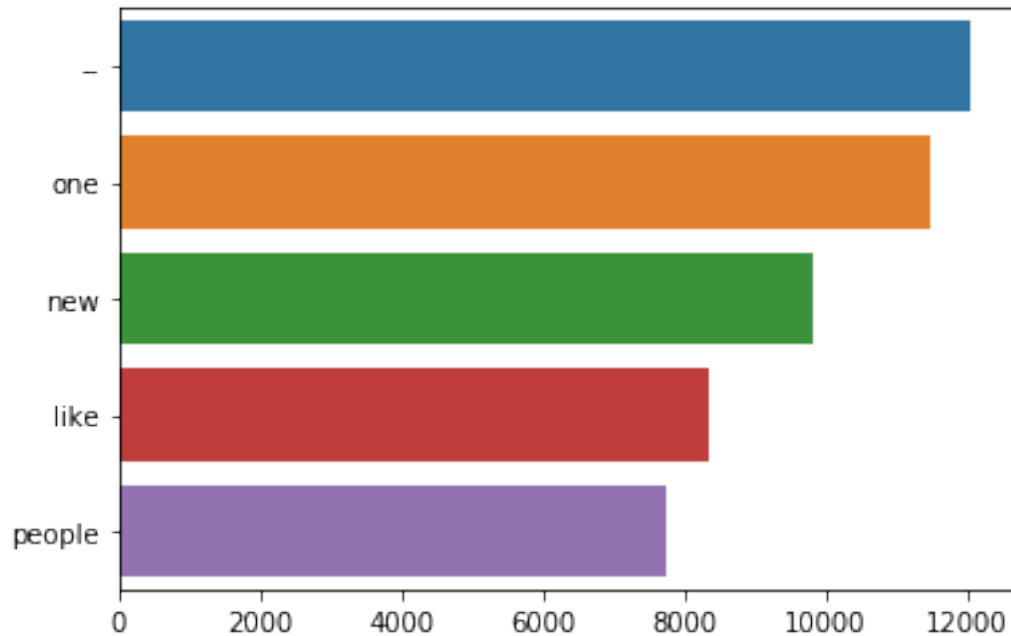
```
[21]: most_common = Counter(corpus).most_common()
```

```
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.

1.2.4 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=axes[line, columna])

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

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



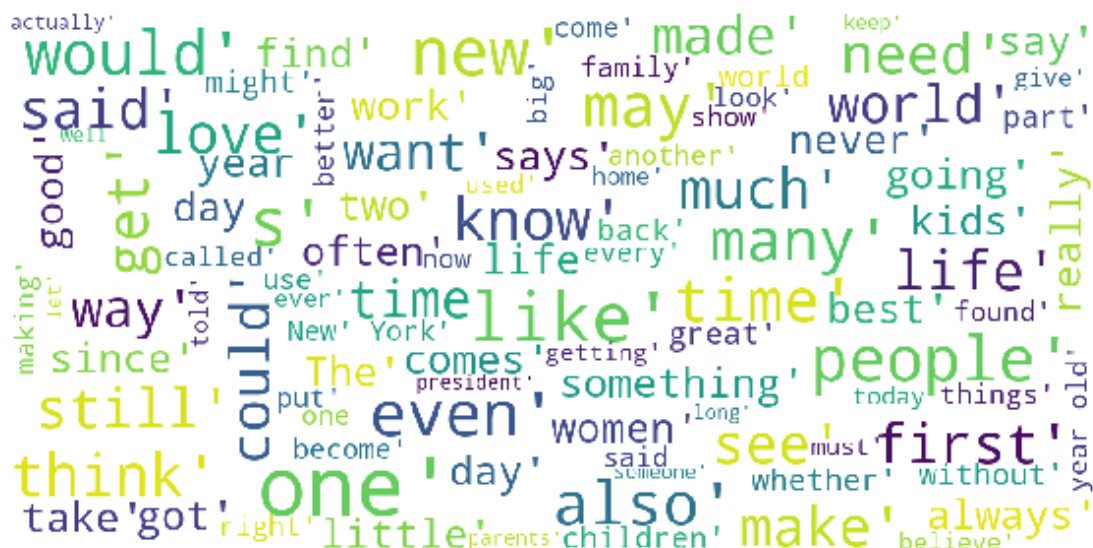
Outra visão importante que podemos notar é o fato de que algumas palavras fazem muito sentido serem mais frequentes em determinadas classes, como alguns exemplos:

- 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.

1.2.5 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.