wordcloud

May 11, 2017

1 Part 1 – WordCloud Creation

The libraries that we are going to use:

```
In [1]: import pandas as pd
        import scipy as sc
        import numpy as np
        import nltk
        import matplotlib.pyplot as plt
        from wordcloud import WordCloud
        from PIL import Image # more beautiful wordcloud the colored.py's based on
        from wordcloud import STOPWORDS, ImageColorGenerator
        from sklearn.feature_extraction import stop_words
```

1.1 Setting up

```
In [2]: # read our training-data
        mydata = pd.read_csv('train_set.csv', sep='\t')
        mydata.head()
```

```
Out[2]:    RowNum     Id                                               Title  \
        0    9560   9561  Sam Adams founder: Beer is more than just 'col...
        1   10801  10802  Slump in oil prices could mean fall in investm...
        2    6726   6727  British Gas owner Centrica warns of higher gas...
        3   12365  12366  Ole Gunnar Solskjaer appointed manager of Card...
        4   11782  11783  Sunderland target loan signings of Kurt Zouma ...

                                                     Content  Category
        0  The craft beer boom, which and been attributed...  Business
        1  The International Energy Agency has warned tha...  Business
        2  Senior executives at British have been accused...  Business
        3   is confident he will have complete control of...  Football
        4  Kurt Zouma and Jack Rodwell are on Sunderland'...  Football
```

We create 5 texts, one for each topic-category, in the form of 5 (very large) strings:

```
In [3]: # create texts
        text_Business = ""
        text_Film = ""
        text_Football = ""
        text_Politics = ""
        text_Technology = ""

        for i in range(mydata.shape[0]):
            if (mydata["Category"][i] == "Business"): text_Business += str(mydata['
            elif(mydata["Category"][i] == "Film"): text_Film += str(mydata["Content
            elif(mydata["Category"][i] == "Football"): text_Football += str(mydata[
            elif(mydata["Category"][i] == "Politics"): text_Politics += str(mydata[
            elif(mydata["Category"][i] == "Technology"): text_Technology += str(my
```

```
In [4]: # our stopwords
        stopwords = set(STOPWORDS) | set(stop_words.ENGLISH_STOP_WORDS)
        # we add whatever additional words we think we need
        stopwords.add('said')
        stopwords.add('say')
        stopwords.add('says')
        stopwords.add('set')
```

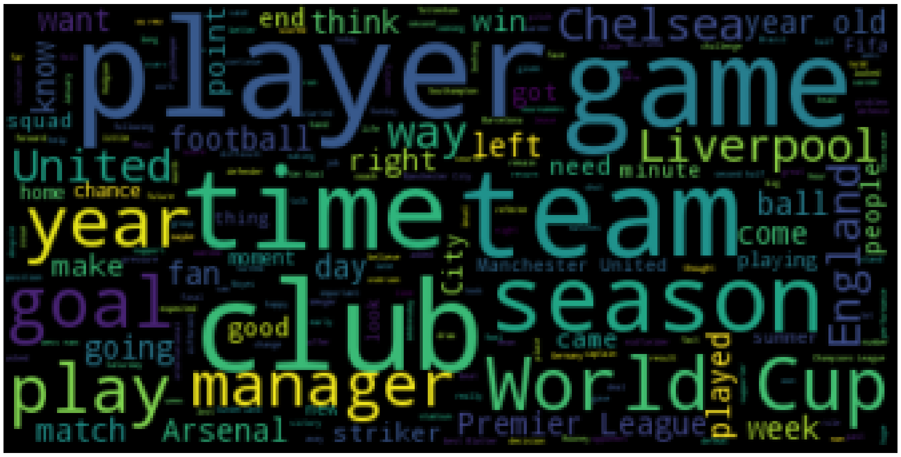
1.2 Creating the WordCloud

The function that we are going to use to create a wordcloud for the argument's *text*, taking into consideration the arguments *stopwords*:

```
In [5]: def create_wordcloud(stopwords, text):
            wc = WordCloud(background_color="black", max_words=200, stopwords=stopw
            # generate the wordcloud as specified
            wc.generate(text)
            plt.imshow(wc, cmap=plt.cm.gray, interpolation="bilinear")
            plt.axis("off")
            plt.show()
```

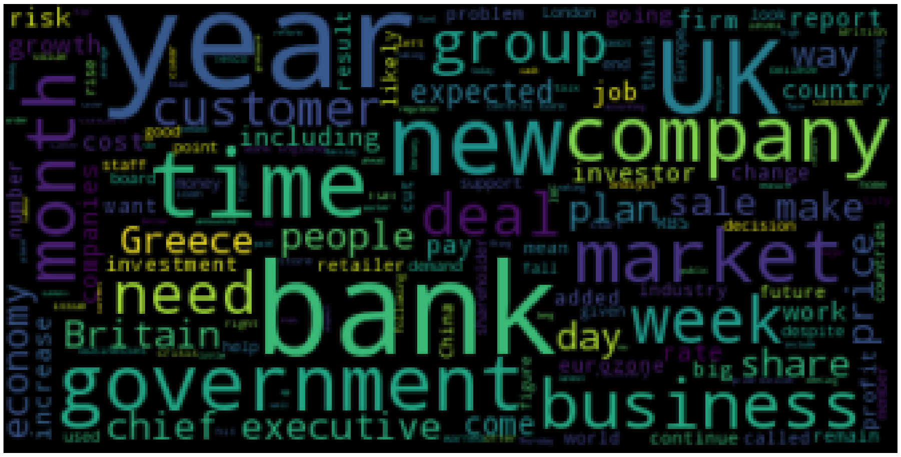
1.2.1 Football's WordCloud

```
In [6]: # create the wordcloud for Football based on the data writen above
        create_wordcloud(stopwords, text_Football)
```



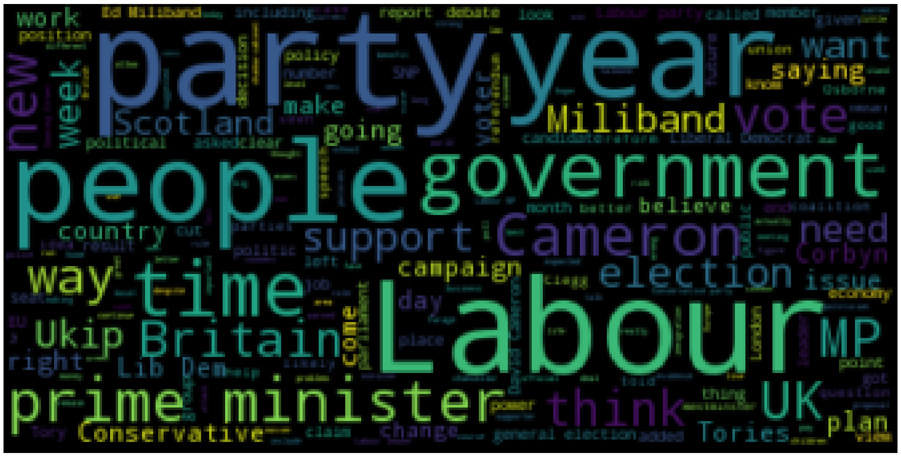
1.2.2 Business' WordCloud

```
In [7]: # create the wordcloud for Business based on the data writen above
        create_wordcloud(stopwords,text_Business)
```



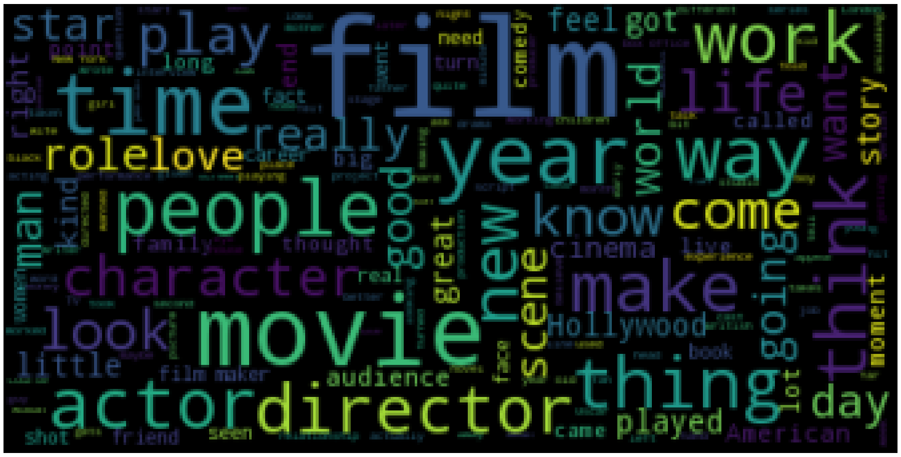
1.2.3 Politics' WordCloud

```
In [8]: # create the wordcloud for Politics based on the data writen above
        create_wordcloud(stopwords, text_Politics)
```



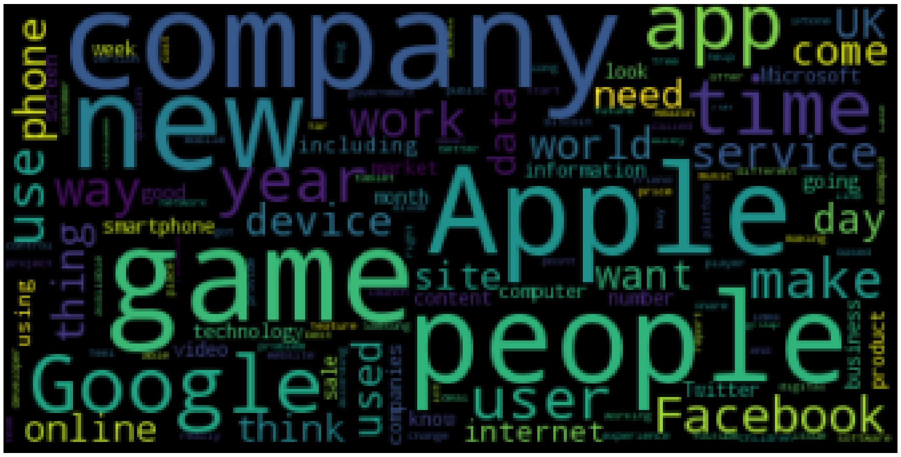
1.2.4 Film's WordCloud

```
In [9]: # create the wordcloud for Film based on the data writen above
        create_wordcloud(stopwords, text_Film)
```



1.2.5 Technology's WordCloud

```
In [10]: # create the wordcloud for Technology based on the data writen above
         create_wordcloud(stopwords,text_Technology)
```


company
new
app
UK
come
time
Microsoft
look
need
phone
use
week
work
data
world
service
way
good
year
Apple
information
going
including
market
month
device
day
thing
smartphone
game
people
site
want
make
using
content
computer
number
technology
video
used
companies
business
product
sale
Google
user
Twitter
Facebook
online
think
know
internet