Extract TEXT from URL

Work developed by Manuel Robalinho at 10/2018

Count the words from a Web Page and present us the words with big influence (the words that occurs more times)

References: https://docs.python.org/3.1/howto/urllib2.html (https://docs.python.org/3.1/howto/urllib2.html)

In [65]:

```
# Libraries
import urllib
from bs4 import BeautifulSoup
#-- Plot
import collections
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
#--- URL Request
import urllib.request
# -- Print
from pprint import pprint
# -- Plot Wordcloud
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from matplotlib import cm
```

In [9]:

```
# Inform the URL
url = "http://news.bbc.co.uk/2/hi/health/2284783.stm"
```

In [10]:

```
# make the request from the URL

req = urllib.request.Request(url)
response = urllib.request.urlopen(req)
the_page = response.read()
```

In [11]:

```
soup = BeautifulSoup(the_page, "lxml")
```

In [12]:

```
# kill all script and style elements
for script in soup(["script", "style"]):
    script.extract()    # rip it out

# get text
text = soup.get_text()

# break into lines and remove leading and trailing space on each
lines = (line.strip() for line in text.splitlines())
# break multi-headlines into a line each
chunks = (phrase.strip() for line in lines for phrase in line.split("  "))
# drop blank lines
text = '\n'.join(chunk for chunk in chunks if chunk)

print(text)
```

BBC NEWS | Health | Blondes 'to die out in 200 years'
NEWS
SPORT
WEATHER
WORLD SERVICE
A-Z INDEX
SEARCH
You are in: Health
News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
Business
Entertainment
Science/Nature
Technology
Health
Medical notes

Talking Point

Country Profiles
In Depth

Programmes

SERVICES
Daily E-mail
News Ticker
Mobile/PDAs

Text Only
Feedback
Help
EDITIONS
Change to UK
Friday, 27 September, 2002, 11:51 GMT 12:51 UK
Blondes 'to die out in 200 years'
Scientists believe the last blondes will be in Finland
The last natural blondes will die out within 200 years, scientists believ
e.
A study by experts in Germany suggests people with blonde hair are an enda
ngered species and will become extinct by 2202.
Researchers predict the last truly natural blonde will be born in Finland
- the country with the highest proportion of blondes.
The frequency of blondes may drop but they won't disappear
Prof Jonathan Rees, University of Edinburgh
But they say too few people now carry the gene for blondes to last beyond
the next two centuries.
The problem is that blonde hair is caused by a recessive gene.
In order for a child to have blonde hair, it must have the gene on both si
des of the family in the grandparents' generation.
Dyed rivals
The researchers also believe that so-called bottle blondes may be to blame
for the demise of their natural rivals.
They suggest that dyed-blondes are more attractive to men who choose them
as partners over true blondes.

Bottle-blondes like Ann Widdecombe may be to blame
But Jonathan Rees, professor of dermatology at the University of Edinburgh said it was unlikely blondes would die out completely.
"Genes don't die out unless there is a disadvantage of having that gene or by chance. They don't disappear," he told BBC News Online.
"The only reason blondes would disappear is if having the gene was a disad vantage and I do not think that is the case.
"The frequency of blondes may drop but they won't disappear."
See also:
28 Mar 01 | Education
What is it about blondes?
09 Apr 99 | Health
Platinum blondes are labelled as dumb
17 Apr 02 | Health
Hair dye cancer alert
Internet links:
University of Edinburgh
The BBC is not responsible for the content of external internet sites
Top Health stories now:
Heart risk link to big families
Back pain drug 'may aid diabetics'
Congo Ebola outbreak confirmed
Vegetables ward off Alzheimer's
Polio campaign launched in Iraq
Gene defect explains high blood pressure
Botox 'may cause new wrinkles'
Alien 'abductees' show real symptoms
Links to more Health stories are at the foot of the page.
E-mail this story to a friend
Links to more Health stories
In This Section
Heart risk link to big families
Back pain drug 'may aid diabetics'
Congo Ebola outbreak confirmed
Vegetables ward off Alzheimer's
Polio campaign launched in Iraq
Gene defect explains high blood pressure
Botox 'may cause new wrinkles'
Alien 'abductees' show real symptoms
How sperm wriggle
Bollywood told to stub it out
Fears over tuna health risk to babies
Public can be taught to spot strokes
^^
Back to top
News Front Page
|
Africa
|
Americas
|
Asia-Pacific
|
Europe
|
Middle East
|
South Asia
|
UK
|

```
Business
|
Entertainment
|
Science/Nature
|
Technology
|
Health
|
Talking Point
|
Country Profiles
|
In Depth
|
Programmes
------------------------------------------------------------------------
--------
To BBC Sport>>
|
To BBC Weather>>
|
To BBC World Service>>
------------------------------------------------------------------------
--------
© MMIII
|
News Sources
|
Privacy
```

In [16]:

```
# print my text file to analyze

pprint(text, width=100)
```

```
("BBC NEWS | Health | Blondes 'to die out in 200 years'\n"
 'NEWS\n'
 'SPORT\n'
 'WEATHER\n'
 'WORLD SERVICE\n'
 'A-Z INDEX\n'
 'SEARCH\n'
 'You are in:\xa0Health\n'
 'News Front Page\n'
 'Africa\n'
 'Americas\n'
 'Asia-Pacific\n'
 'Europe\n'
 'Middle East\n'
 'South Asia\n'
 'UK\n'
 'Business\n'
 'Entertainment\n'
 'Science/Nature\n'
 'Technology\n'
 'Health\n'
 'Medical notes\n'
 '-------------\n'
 'Talking Point\n'
 '-------------\n'
 'Country Profiles\n'
 'In Depth\n'
 '-------------\n'
 'Programmes\n'
 '-------------\n'
 'SERVICES\n'
 'Daily E-mail\n'
 'News Ticker\n'
 'Mobile/PDAs\n'
 '-------------\n'
 'Text Only\n'
 'Feedback\n'
 'Help\n'
 'EDITIONS\n'
 'Change to UK\n'
 'Friday, 27 September, 2002, 11:51 GMT 12:51 UK\n'
 "Blondes 'to die out in 200 years'\n"
 'Scientists believe the last blondes will be in Finland\n'
 'The last natural blondes will die out within 200 years, scientists belie
ve.\n'
 'A study by experts in Germany suggests people with blonde hair are an en
dangered species and '
 'will become extinct by 2202.\n'
 'Researchers predict the last truly natural blonde will be born in Finlan
d - the country with the '
 'highest proportion of blondes.\n'
 "The frequency of blondes may drop but they won't disappear\n"
 'Prof Jonathan Rees, University of Edinburgh\n'
 'But they say too few people now carry the gene for blondes to last beyon
d the next two '
 'centuries.\n'
 'The problem is that blonde hair is caused by a recessive gene.\n'
 'In order for a child to have blonde hair, it must have the gene on both
sides of the family in '
 "the grandparents' generation.\n"
 'Dyed rivals\n'
```

```
 'The researchers also believe that so-called bottle blondes may be to bla
me for the demise of '
 'their natural rivals.\n'
 'They suggest that dyed-blondes are more attractive to men who choose the
m as partners over true '
 'blondes.\n'
 'Bottle-blondes like Ann Widdecombe may be to blame\n'
 'But Jonathan Rees, professor of dermatology at the University of Edinbur
gh said it was unlikely '
 'blondes would die out completely.\n'
 '"Genes don\'t die out unless there is a disadvantage of having that gene
or by chance. They '
 'don\'t disappear," he told BBC News Online.\n'
 '"The only reason blondes would disappear is if having the gene was a dis
advantage and I do not '
 'think that is the case.\n'
 '"The frequency of blondes may drop but they won\'t disappear."\n'
 'See also:\n'
 '28 Mar 01\xa0|\xa0Education\n'
 'What is it about blondes?\n'
 '09 Apr 99\xa0|\xa0Health\n'
 'Platinum blondes are labelled as dumb\n'
 '17 Apr 02\xa0|\xa0Health\n'
 'Hair dye cancer alert\n'
 'Internet links:\n'
 'University of Edinburgh\n'
 'The BBC is not responsible for the content of external internet sites\n'
 'Top Health stories now:\n'
 'Heart risk link to big families\n'
 "Back pain drug 'may aid diabetics'\n"
 'Congo Ebola outbreak confirmed\n'
 "Vegetables ward off Alzheimer's\n"
 'Polio campaign launched in Iraq\n'
 'Gene defect explains high blood pressure\n'
 "Botox 'may cause new wrinkles'\n"
 "Alien 'abductees' show real symptoms\n"
 'Links to more Health stories are at the foot of the page.\n'
 'E-mail this story to a friend\n'
 'Links to more Health stories\n'
 'In This Section\n'
 'Heart risk link to big families\n'
 "Back pain drug 'may aid diabetics'\n"
 'Congo Ebola outbreak confirmed\n'
 "Vegetables ward off Alzheimer's\n"
 'Polio campaign launched in Iraq\n'
 'Gene defect explains high blood pressure\n'
 "Botox 'may cause new wrinkles'\n"
 "Alien 'abductees' show real symptoms\n"
 'How sperm wriggle\n'
 'Bollywood told to stub it out\n'
 'Fears over tuna health risk to babies\n'
 'Public can be taught to spot strokes\n'
 '^^\n'
 'Back to top\n'
 'News Front Page\n'
 '|\n'
 'Africa\n'
 '|\n'
 'Americas\n'
 '|\n'
 'Asia-Pacific\n'
```

```
 '|\n'
 'Europe\n'
 '|\n'
 'Middle East\n'
 '|\n'
 'South Asia\n'
 '|\n'
 'UK\n'
 '|\n'
 'Business\n'
 '|\n'
 'Entertainment\n'
 '|\n'
 'Science/Nature\n'
 '|\n'
 'Technology\n'
 '|\n'
 'Health\n'
 '|\n'
 'Talking Point\n'
 '|\n'
 'Country Profiles\n'
 '|\n'
 'In Depth\n'
 '|\n'
 'Programmes\n'
 '------------------------------------------------------------------------
----------\n'
 'To BBC Sport>>\n'
 '|\n'
 'To BBC Weather>>\n'
 '|\n'
 'To BBC World Service>>\n'
 '------------------------------------------------------------------------
----------\n'
 '© MMIII\n'
 '|\n'
 'News Sources\n'
 '|\n'
 'Privacy')
```

In [21]:

```
# Path to acess the file
path = 'ml/count_words/'
```

In [53]:

```
# Stopwords - File with words we don't need count
# read one word each line of stopwords
stopwords = set(line.strip() for line in open(path + 'stopwords_en.txt'))
stopwords = stopwords.union(set(['mr','mrs','one','two','said']))
```

In [54]:

```
# Print the words i dont't want to count ( stopword )
print (stopwords)
```

```
{'', 'a', 'new', 'one', 'as', 'see', 'true', 'them', '7', 'to', 'their',
'a-z', 'are', 'was', "won't", 'the', 'off', 'be', 'show', 'ann', '9', 'i
s', 'mar', 'of', 'what', '02', 'an', 'there', '4', 'am', 'next', '09',
'1', 'that', 'mrs', '6', 'more', '01', '04', 'by', '10', 'he', 'other', 'n
otes', 'than', '3', 'you', 'two', 'page', 'on', 'i', 'with', 'your', '06',
"don't", '^^', 'can', 'they', '08', 'text', 'gmt', 'for', '8', 'last',
'5', 'only', 'who', 'top', 'thing', 'say', 'not', 'it', 'over', 'said', 'n
ews', '03', 'may', 'out', 'back', '07', '05', 'help', '2', 'but', 'will',
'while', 'how', 'in', 'within', 'these', 'mr', 'too', 'or', 'all', 'if',
'0', 'do', 'index', 'and', 'this', 'at'}
```

In [55]:

```
# Instantiate a dictionary, and for every word in the file,
# Add to the dictionary if it doesn't exist. If it does, increase the count.
wordcount = {}
```

In [56]:

```
# To eliminate duplicates, split by punctuation, and use case demiliters.
text_file = text

for word in text_file.lower().split():
    word = word.replace(".","")
    word = word.replace(",","")
    word = word.replace(":","")
    word = word.replace("\"","")
    word = word.replace("!","")
    word = word.replace("--","")
    word = word.replace("|","")
    word = word.replace("â€œ","")
    word = word.replace("â€˜","")
    word = word.replace("*","")
    word = word.replace("©","")
    if word not in stopwords:
        if word not in wordcount:
            wordcount[word] = 1
        else:
            wordcount[word] += 1
```

In [57]:

```
pprint(wordcount)
```

```
{"'abductees'": 2,
 "'may": 4,
 "'to": 2,
 '-': 6,
 '1151': 1,
 '1251': 1,
 '17': 1,
 '200': 3,
 '2002': 1,
 '2202': 1,
 '27': 1,
 '28': 1,
 '99': 1,
 'about': 1,
 'africa': 2,
 'aid': 2,
 'alert': 1,
 'alien': 2,
 'also': 2,
 "alzheimer's": 2,
 'americas': 2,
 'apr': 2,
 'asia': 2,
 'asia-pacific': 2,
 'attractive': 1,
 'babies': 1,
 'bbc': 6,
 'become': 1,
 'believe': 3,
 'beyond': 1,
 'big': 2,
 'blame': 2,
 'blonde': 4,
 'blondes': 13,
 'blondes?': 1,
 'blood': 2,
 'bollywood': 1,
 'born': 1,
 'both': 1,
 'botox': 2,
 'bottle': 1,
 'bottle-blondes': 1,
 'business': 2,
 'campaign': 2,
 'cancer': 1,
 'carry': 1,
 'case': 1,
 'cause': 2,
 'caused': 1,
 'centuries': 1,
 'chance': 1,
 'change': 1,
 'child': 1,
 'choose': 1,
 'completely': 1,
 'confirmed': 2,
 'congo': 2,
 'content': 1,
 'country': 3,
 'daily': 1,
 'defect': 2,
```

```
    'demise': 1,
    'depth': 2,
    'dermatology': 1,
    "diabetics'": 2,
    'die': 5,
    'disadvantage': 2,
    'disappear': 4,
    'drop': 2,
    'drug': 2,
    'dumb': 1,
    'dye': 1,
    'dyed': 1,
    'dyed-blondes': 1,
    'e-mail': 2,
    'east': 2,
    'ebola': 2,
    'edinburgh': 3,
    'editions': 1,
    'education': 1,
    'endangered': 1,
    'entertainment': 2,
    'europe': 2,
    'experts': 1,
    'explains': 2,
    'external': 1,
    'extinct': 1,
    'families': 2,
    'family': 1,
    'fears': 1,
    'feedback': 1,
    'few': 1,
    'finland': 2,
    'foot': 1,
    'frequency': 2,
    'friday': 1,
    'friend': 1,
    'front': 2,
    'gene': 7,
    'generation': 1,
    'genes': 1,
    'germany': 1,
    "grandparents'": 1,
    'hair': 4,
    'have': 2,
    'having': 2,
    'health': 10,
    'heart': 2,
    'high': 2,
    'highest': 1,
    'internet': 2,
    'iraq': 2,
    'jonathan': 2,
    'labelled': 1,
    'launched': 2,
    'like': 1,
    'link': 2,
    'links': 3,
    'medical': 1,
    'men': 1,
    'middle': 2,
    'mmiii': 1,
```

```
    'mobile/pdas': 1,
    'must': 1,
    'natural': 3,
    'now': 2,
    'online': 1,
    'order': 1,
    'outbreak': 2,
    'pain': 2,
    'partners': 1,
    'people': 2,
    'platinum': 1,
    'point': 2,
    'polio': 2,
    'predict': 1,
    'pressure': 2,
    'privacy': 1,
    'problem': 1,
    'prof': 1,
    'professor': 1,
    'profiles': 2,
    'programmes': 2,
    'proportion': 1,
    'public': 1,
    'real': 2,
    'reason': 1,
    'recessive': 1,
    'rees': 2,
    'researchers': 2,
    'responsible': 1,
    'risk': 3,
    'rivals': 2,
    'science/nature': 2,
    'scientists': 2,
    'search': 1,
    'section': 1,
    'september': 1,
    'service': 1,
    'service>>': 1,
    'services': 1,
    'sides': 1,
    'sites': 1,
    'so-called': 1,
    'sources': 1,
    'south': 2,
    'species': 1,
    'sperm': 1,
    'sport': 1,
    'sport>>': 1,
    'spot': 1,
    'stories': 3,
    'story': 1,
    'strokes': 1,
    'stub': 1,
    'study': 1,
    'suggest': 1,
    'suggests': 1,
    'symptoms': 2,
    'talking': 2,
    'taught': 1,
    'technology': 2,
    'think': 1,
```

```
 'ticker': 1,
 'told': 2,
 'truly': 1,
 'tuna': 1,
 'uk': 4,
 'university': 3,
 'unless': 1,
 'unlikely': 1,
 'vegetables': 2,
 'ward': 2,
 'weather': 1,
 'weather>>': 1,
 'widdecombe': 1,
 'world': 2,
 'would': 2,
 'wriggle': 1,
 "wrinkles'": 2,
 'years': 1,
 "years'": 2}
```

In [ ]:

In [59]:

```
# Print most common word
n_print = int(input("How many most common words to print: "))
print("\nOK. The {} most common words are as follows\n".format(n_print))
word_counter = collections.Counter(wordcount)
for word, count in word_counter.most_common(n_print):
    print(word, ": ", count)
```

```
How many most common words to print: 30

OK. The 30 most common words are as follows

blondes :  13
health :  10
gene :  7
bbc :  6
- :  6
die :  5
uk :  4
blonde :  4
hair :  4
disappear :  4
'may :  4
200 :  3
country :  3
believe :  3
natural :  3
university :  3
edinburgh :  3
links :  3
stories :  3
risk :  3
'to :  2
years' :  2
world :  2
front :  2
africa :  2
americas :  2
asia-pacific :  2
europe :  2
middle :  2
east :  2
```

In [70]:

```
# Transform in a Data Frame
lst = word_counter.most_common(n_print)
df = pd.DataFrame(lst, columns = ['Word', 'Count'])

# Select first registers
df1 = df.head(30)   # Put the number you want
df1.head(10)
```

Out[70]:

	Word	Count
0	blondes	13
1	health	10
2	gene	7
3	bbc	6
4	-	6
5	die	5
6	uk	4
7	blonde	4
8	hair	4
9	disappear	4

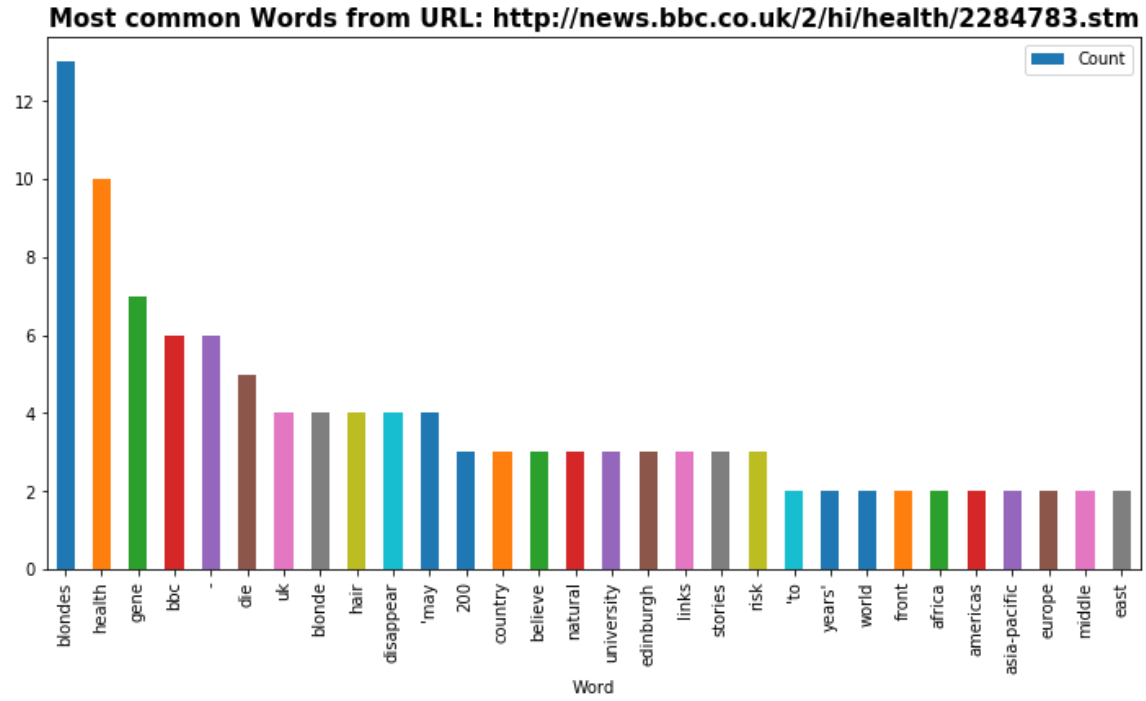
In [71]:

```
my_title = "Most common Words from URL: "+url
```

In [72]:

```
# Create a data frame of the most common words
# Draw a bar chart

fig, ax1 = plt.subplots(figsize=(12,6))
ax1 = df1.plot.bar(ax=ax1, x='Word',y='Count')
plt.title(my_title , fontdict={'size':15, 'weight': 'bold'});
```



In [73]:

```
wrds = df.Word

# WORDS without spaces
wrds =  df1["Word"].str.replace(" ","")
wrds.head()
```

Out[73]:

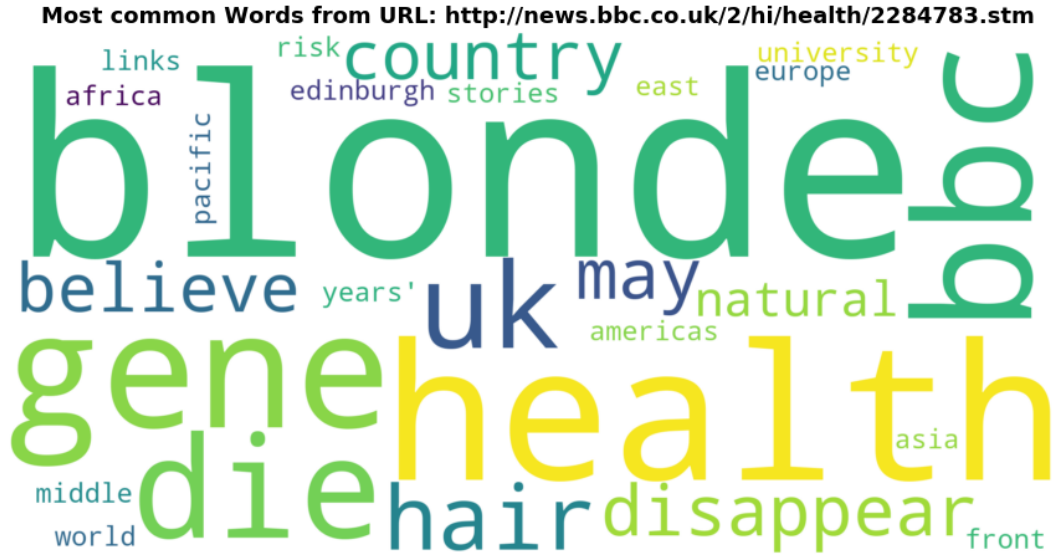
```
0    blondes
1     health
2       gene
3        bbc
4          -
Name: Word, dtype: object
```

In [74]:

```
# Plot The WORDS in a Frame

wc = WordCloud( background_color='white', colormap=cm.viridis, scale=5).generate(" ".jo
in(wrds))

plt.figure(figsize=(16,8))
plt.imshow(wc, interpolation="bilinear", origin='upper')
plt.axis("off")
plt.title(my_title , fontdict={'size':18, 'weight': 'bold'});
```



In [ ]: