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 [1]:

```
# 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 [2]:

```
# Inform the URL
#url = "http://news.bbc.co.uk/2/hi/health/2284783.stm"
url = "https://www.bbc.com/news/technology-45747983.stm"
```

In [3]:

```
# make the request from the URL

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

In [4]:

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

In [5]:

```
# 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)
```

```
'China spy attack hits Apple and Amazon' - BBC News
HomepageAccessibility linksSkip to contentAccessibility HelpBBC iDNotifica
tionsHomeNewsSportWeatheriPlayerTVRadioCBBCCBeebiesFoodBitesizeMusicEarthA
rtsMake It DigitalTasterLocalTomorrow's WorldMenuSearchSearch the BBCSearc
h the BBC
News
BBC News Navigation
Sections
Home
Video
World
UK
Business
Tech
selected
Science
Stories
Entertainment & Arts
Health
World News TV
In Pictures
Reality Check
Newsbeat
Special Reports
Explainers
The Reporters
Have Your Say
Technology
Technology
'China spy attack hits Apple and Amazon'
4 October 2018
Share this with Facebook
Share this with Messenger
Share this with Twitter
Share this with Email
Share this with Facebook
Share this with WhatsApp
Share this with Messenger
Share this with Twitter
Share
Share this with
These are external links and will open in a new window
Email
Share this with Email
Facebook
Share this with Facebook
Messenger
Share this with Messenger
Messenger
Share this with Messenger
Twitter
Share this with Twitter
Pinterest
Share this with Pinterest
WhatsApp
Share this with WhatsApp
LinkedIn
Share this with LinkedIn
Copy this link
https://www.bbc.com/news/technology-45747983.stm
Read more about sharing.
```

```
These are external links and will open in a new window
Close share panel
Image copyright
AFP
Image caption
US warships were found to be harbouring the compromised computers, Bloombe
rg says
Apple and Amazon are among US companies and agencies who have had data sto
len by Chinese spies, claims Bloomberg.The data had been siphoned off via 
tiny chips inserted on server circuit boards made by a company called Supe
r Micro Computer, reported the news agency.The servers had been compromise
d during manufacturing and the chips activated once they were up and runni
ng, it said.Apple, Amazon and Super Micro have rejected Bloomberg's claim
s, calling them "untrue".In particular, Apple released a strong statement 
in response to Bloomberg's article saying it had found "no evidence" to su
pport the allegations.
Bloomberg said a year-long investigation by reporters Jordan Robertson and
Michael Riley had uncovered evidence of the wide-ranging attack, which gav
e Beijing access to 30 large companies and many federal agencies.
US warns of supply chain cyber-attacks
Pentagon warns on compromised code
Trump relaxes rules around cyber-attacks
It said the first information about the spying campaign had emerged during
security testing carried out by Amazon in 2015 before it had started using
servers from US company Elemental, which had been manufactured by Super Mi
cro Computer at plants in China. And this discovery then kicked off a long
-running "top-secret probe" by US intelligence agencies, which found compr
omised servers:
in Department of Defense data centres
onboard warships
handling data gathered by CIA drones
China was well placed to carry out this kind of attack, said Bloomberg, be
cause 90% of the world's PCs are made in the country. Carrying out the att
ack involved "developing a deep understanding of a product's design, manip
ulating components at the factory, and ensuring that the doctored devices 
made it through the global logistics chain to the desired location", it sa
id.
Image copyright
Getty Images
Image caption
Many companies have been caught out by software maliciously modified befor
e it reaches them
Many US companies, including Apple, Amazon and major banks, were also usin
g Super Micro Computer hardware. Bloomberg claims the probe led to some co
mpanies removing servers made by Super Micro and ending business relations
hips with the company.Amazon and Apple both denied there was any substance
to Bloomberg's claims.In its lengthy statement, Amazon said: "We've found 
no evidence to support claims of malicious chips or hardware modification
s."Apple took Bloomberg to task, saying the agency had contacted it "multi
ple times with claims, sometimes vague and sometimes elaborate, of an alle
ged security incident"."Each time, we have conducted rigorous internal inv
estigations based on their inquiries and each time we have found absolutel
y no evidence to support any of them."It added: "We have repeatedly and co
nsistently offered factual responses, on the record, refuting virtually ev
ery aspect of Bloomberg's story relating to Apple."Super Micro Computer sa
id it was "not aware" of any government investigation into the issue and n
o customer had stopped using its products because of fears about Chinese h
ackers.China's Ministry of Foreign Affairs called the story a "gratuitous 
accusation" and said the safety of supply chains was an "issue of common c
oncern".Bloomberg said the denials were countered by testimony from "six c
urrent and former national security officials" as well as insiders at both
```

```
Apple and Amazon who had detailed the investigation and its aftermath.
Related TopicsChinaCyber-attackAmazonApple
Share this story
About sharing
Email
Facebook
Messenger
Messenger
Twitter
Pinterest
WhatsApp
LinkedIn
More on this story
US warns of supply chain cyber-attacks
26 July 2018
British Airways breach: How did hackers get in?
7 September 2018
President Trump relaxes US cyber-attacks rules
16 August 2018
US military draws up 'do not buy' list for software
30 July 2018
Top Stories
Brazil exit polls give Bolsonaro wide lead
A far-right presidential candidate looks set to go through to a second rou
nd, exit polls suggest.
7 October 2018
China confirms Interpol chief detained
7 October 2018
Saudi writer 'murdered' in consulate
7 October 2018
Features
Murder mystery: The Reykjavik Confessions
Video
My life as Pablo Escobar's lovechild
Can chocolate tasting unite Trump’s America?
Video
'Why I've bought a ticket to the Moon'
Fighting for the right to be a prostitute
BBC Travel: The town that throws wine in the sea
Pop charts... as you've never seen them
Do you chop your onions? Food blogger meets Michelin chef
'I'm no longer ashamed of my disabled daughter'
Elsewhere on the BBC
Lyrics quiz
Have you been getting these songs wrong?
Full article Lyrics quiz
Feeling hot
What happens to your body in extreme heat?
Full article Feeling hot
Why you can trust BBC News
BBC News Navigation
Sections
Home
Video
World
World Home
Africa
Asia
Australia
Europe
Latin America
```

```
Middle East
US & Canada
UK
UK Home
England
N. Ireland
Scotland
Wales
Politics
Business
Business Home
Market Data
Global Trade
Companies
Entrepreneurship
Technology of Business
Connected World
Global Education
Economy
Tech
selected
Science
Stories
Entertainment & Arts
Health
World News TV
In Pictures
Reality Check
Newsbeat
Special Reports
Explainers
The Reporters
Have Your Say
BBC News Services
On your mobile
On your connected tv
Get news alerts
Contact BBC News
Explore the BBCHomeNewsSportWeatheriPlayerTVRadioCBBCCBeebiesFoodBitesizeM
usicEarthArtsMake It DigitalTasterLocalTomorrow's WorldTerms of UseAbout t
he BBCPrivacy PolicyCookiesAccessibility HelpParental GuidanceContact the
BBCGet Personalised NewslettersCopyright © 2018 BBC. The BBC is not respon
sible for the content of external sites. Read about our approach to extern
al linking.
```

In [6]:

```
# print my text file to analyze

pprint(text, width=100)
```

```
("'China spy attack hits Apple and Amazon' - BBC News\n"
 'HomepageAccessibility linksSkip to contentAccessibility HelpBBC '
 'iDNotificationsHomeNewsSportWeatheriPlayerTVRadioCBBCCBeebiesFoodBitesiz
eMusicEarthArtsMake It '
 "DigitalTasterLocalTomorrow's WorldMenuSearchSearch the BBCSearch the BBC
\n"
 'News\n'
 'BBC News Navigation\n'
 'Sections\n'
 'Home\n'
 'Video\n'
 'World\n'
 'UK\n'
 'Business\n'
 'Tech\n'
 'selected\n'
 'Science\n'
 'Stories\n'
 'Entertainment & Arts\n'
 'Health\n'
 'World News TV\n'
 'In Pictures\n'
 'Reality Check\n'
 'Newsbeat\n'
 'Special Reports\n'
 'Explainers\n'
 'The Reporters\n'
 'Have Your Say\n'
 'Technology\n'
 'Technology\n'
 "'China spy attack hits Apple and Amazon'\n"
 '4 October 2018\n'
 'Share this with Facebook\n'
 'Share this with Messenger\n'
 'Share this with Twitter\n'
 'Share this with Email\n'
 'Share this with Facebook\n'
 'Share this with WhatsApp\n'
 'Share this with Messenger\n'
 'Share this with Twitter\n'
 'Share\n'
 'Share this with\n'
 'These are external links and will open in a new window\n'
 'Email\n'
 'Share this with Email\n'
 'Facebook\n'
 'Share this with Facebook\n'
 'Messenger\n'
 'Share this with Messenger\n'
 'Messenger\n'
 'Share this with Messenger\n'
 'Twitter\n'
 'Share this with Twitter\n'
 'Pinterest\n'
 'Share this with Pinterest\n'
 'WhatsApp\n'
 'Share this with WhatsApp\n'
 'LinkedIn\n'
 'Share this with LinkedIn\n'
 'Copy this link\n'
 'https://www.bbc.com/news/technology-45747983.stm\n'
```

```
  'Read more about sharing.\n'
  'These are external links and will open in a new window\n'
  'Close share panel\n'
  'Image copyright\n'
  'AFP\n'
  'Image caption\n'
  'US warships were found to be harbouring the compromised computers, Bloom
 berg says\n'
  'Apple and Amazon are among US companies and agencies who have had data s
 tolen by Chinese spies, '
  'claims Bloomberg.The data had been siphoned off via tiny chips inserted 
 on server circuit boards '
  'made by a company called Super Micro Computer, reported the news agency.
 The servers had been '
  'compromised during manufacturing and the chips activated once they were 
 up and running, it '
  'said.Apple, Amazon and Super Micro have rejected Bloomberg\'s claims, ca
 lling them "untrue".In '
  "particular, Apple released a strong statement in response to Bloomberg's
 article saying it had "
  'found "no evidence" to support the allegations.\n'
  'Bloomberg said a year-long investigation by reporters Jordan Robertson a
 nd Michael Riley had '
  'uncovered evidence of the wide-ranging attack, which gave Beijing access
 to 30 large companies '
  'and many federal agencies.\n'
  'US warns of supply chain cyber-attacks\n'
  'Pentagon warns on compromised code\n'
  'Trump relaxes rules around cyber-attacks\n'
  'It said the first information about the spying campaign had emerged duri
 ng security testing '
  'carried out by Amazon in 2015 before it had started using servers from U
 S company Elemental, '
  'which had been manufactured by Super Micro Computer at plants in China. 
 And this discovery then '
  'kicked off a long-running "top-secret probe" by US intelligence agencie
 s, which found '
  'compromised servers:\n'
  'in Department of Defense data centres\n'
  'onboard warships\n'
  'handling data gathered by CIA drones\n'
  'China was well placed to carry out this kind of attack, said Bloomberg, 
 because 90% of the '
  'world\'s PCs are made in the country. Carrying out the attack involved 
 "developing a deep '
  "understanding of a product's design, manipulating components at the fact
 ory, and ensuring that "
  'the doctored devices made it through the global logistics chain to the d
 esired location", it '
  'said.\n'
  'Image copyright\n'
  'Getty Images\n'
  'Image caption\n'
  'Many companies have been caught out by software maliciously modified bef
 ore it reaches them\n'
  'Many US companies, including Apple, Amazon and major banks, were also us
 ing Super Micro Computer '
  'hardware. Bloomberg claims the probe led to some companies removing serv
 ers made by Super Micro '
  'and ending business relationships with the company.Amazon and Apple both
 denied there was any '
  'substance to Bloomberg\'s claims.In its lengthy statement, Amazon said:
 "We\'ve found no '
  'evidence to support claims of malicious chips or hardware modification
 s."Apple took Bloomberg to '
  'task, saying the agency had contacted it "multiple times with claims, so
 metimes vague and '
  'sometimes elaborate, of an alleged security incident"."Each time, we hav
 e conducted rigorous '
  'internal investigations based on their inquiries and each time we have f
 ound absolutely no '
  'evidence to support any of them."It added: "We have repeatedly and consi
 stently offered factual '
  "responses, on the record, refuting virtually every aspect of Bloomberg's
 story relating to "
  'Apple."Super Micro Computer said it was "not aware" of any government in
 vestigation into the '
  'issue and no customer had stopped using its products because of fears ab
 out Chinese '
  'hackers.China\'s Ministry of Foreign Affairs called the story a "gratuit
 ous accusation" and said '
  'the safety of supply chains was an "issue of common concern".Bloomberg s
 aid the denials were '
  'countered by testimony from "six current and former national security of
 ficials" as well as '
  'insiders at both Apple and Amazon who had detailed the investigation and
 its aftermath.\n'
  'Related TopicsChinaCyber-attackAmazonApple\n'
  'Share this story\n'
  'About\xa0sharing\n'
  'Email\n'
  'Facebook\n'
  'Messenger\n'
  'Messenger\n'
  'Twitter\n'
  'Pinterest\n'
  'WhatsApp\n'
  'LinkedIn\n'
  'More on this story\n'
  'US warns of supply chain cyber-attacks\n'
  '26 July 2018\n'
  'British Airways breach: How did hackers get in?\n'
  '7 September 2018\n'
  'President Trump relaxes US cyber-attacks rules\n'
  '16 August 2018\n'
  "US military draws up 'do not buy' list for software\n"
  '30 July 2018\n'
  'Top Stories\n'
  'Brazil exit polls give Bolsonaro wide lead\n'
  'A far-right presidential candidate looks set to go through to a second r
 ound, exit polls '
  'suggest.\n'
  '7 October 2018\n'
  'China confirms Interpol chief detained\n'
  '7 October 2018\n'
  "Saudi writer 'murdered' in consulate\n"
  '7 October 2018\n'
  'Features\n'
  'Murder mystery: The Reykjavik Confessions\n'
  'Video\n'
  "My life as Pablo Escobar's lovechild\n"
  'Can chocolate tasting unite Trump’s America?\n'
```

```
    'Video\n'
    "'Why I've bought a ticket to the Moon'\n"
    'Fighting for the right to be a prostitute\n'
    'BBC Travel: The town that throws wine in the sea\n'
    "Pop charts... as you've never seen them\n"
    'Do you chop your onions? Food blogger meets Michelin chef\n'
    "'I'm no longer ashamed of my disabled daughter'\n"
    'Elsewhere on the BBC\n'
    'Lyrics quiz\n'
    'Have you been getting these songs wrong?\n'
    'Full article Lyrics quiz\n'
    'Feeling hot\n'
    'What happens to your body in extreme heat?\n'
    'Full article Feeling hot\n'
    'Why you can trust BBC News\n'
    'BBC News Navigation\n'
    'Sections\n'
    'Home\n'
    'Video\n'
    'World\n'
    'World Home\n'
    'Africa\n'
    'Asia\n'
    'Australia\n'
    'Europe\n'
    'Latin America\n'
    'Middle East\n'
    'US & Canada\n'
    'UK\n'
    'UK Home\n'
    'England\n'
    'N. Ireland\n'
    'Scotland\n'
    'Wales\n'
    'Politics\n'
    'Business\n'
    'Business Home\n'
    'Market Data\n'
    'Global Trade\n'
    'Companies\n'
    'Entrepreneurship\n'
    'Technology of Business\n'
    'Connected World\n'
    'Global Education\n'
    'Economy\n'
    'Tech\n'
    'selected\n'
    'Science\n'
    'Stories\n'
    'Entertainment & Arts\n'
    'Health\n'
    'World News TV\n'
    'In Pictures\n'
    'Reality Check\n'
    'Newsbeat\n'
    'Special Reports\n'
    'Explainers\n'
    'The Reporters\n'
    'Have Your Say\n'
    'BBC News Services\n'
    'On your mobile\n'
```

```
 'On your connected tv\n'
 'Get news alerts\n'
 'Contact BBC News\n'
 'Explore the BBCHomeNewsSportWeatheriPlayerTVRadioCBBCCBeebiesFoodBitesiz
eMusicEarthArtsMake It '
 "DigitalTasterLocalTomorrow's WorldTerms of UseAbout the BBCPrivacy Polic
yCookiesAccessibility "
 'HelpParental GuidanceContact the BBCGet Personalised NewslettersCopyrigh
t © 2018 BBC. The BBC is '
 'not responsible for the content of external sites. Read about our approa
ch to external linking.')
```

In [7]:

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

In [19]:

```
# 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 [20]:

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

```
{'at', 'or', 'these', 'a-z', 'your', 'while', 'mr', 'two', 'longer', 'emai
l', 'one', 'but', 'of', 'a', 'new', '08', 'more', 'gmt', 'thing', '8', 'i
s', 'who', '26', '4', 'am', '2018', 'you', '0', 'may', 'page', 'on', 'int
o', 'an', 'back', 'made', '6', '3', 'by', 'hot', 'why', 'next', 'home', 's
ee', 'index', 'this', "don't", '9', '16', '05', 'as', 'both', 'other',
'5', 'text', 'he', 'us', 'large', 'found', 'over', '07', 'top', 'true', 's
aid', '1', 'say', 'have', 'did', 'within', '01', 'there', '06', 'they', 'l
ist', 'how', 'many', 'former', '09', 'every', 'too', '10', '7', 'added',
'july', 'says', 'was', '30', 'to', '2', 'which', "won't", 'link', 'the',
'their', 'exit', 'last', 'them', 'around', 'for', 'notes', '02', '03', 'he
lp', 'first', 'its', 'had', 'are', 'close', 'i', 'that', 'all', 'also', 's
how', 'do', 'what', '90%', 'and', 'it', 'mar', 'if', 'will', 'than', 'mr
s', 'about', 'only', '2015', 'news', 'ann', 'no', 'out', 'can', 'be', 'no
t', 'been', 'in', 'any', 'with', 'off', '04'}
```

In [21]:

```
# 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 [22]:

```
# 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 [23]:

```
pprint(wordcount)
```

```
{'': 1,
 '&': 3,
 "'china": 2,
 "'do": 1,
 "'i'm": 1,
 "'murdered'": 1,
 "'why": 1,
 '-': 1,
 'absolutely': 1,
 'access': 1,
 'accusation': 1,
 'activated': 1,
 'affairs': 1,
 'afp': 1,
 'africa': 1,
 'aftermath': 1,
 'agencies': 3,
 'agency': 1,
 'agencythe': 1,
 'airways': 1,
 'alerts': 1,
 'allegations': 1,
 'alleged': 1,
 'amazon': 6,
 "amazon'": 2,
 'america': 1,
 'america?': 1,
 'among': 1,
 'apple': 7,
 'applesuper': 1,
 'approach': 1,
 'article': 3,
 'arts': 2,
 'ashamed': 1,
 'asia': 1,
 'aspect': 1,
 'attack': 5,
 'august': 1,
 'australia': 1,
 'aware': 1,
 'banks': 1,
 'based': 1,
 'bbc': 11,
 'bbcget': 1,
 'bbchomenewssportweatheriplayertvradiocbbccbeebiesfoodbitesizemusicearthа
rtsmake': 1,
 'bbcprivacy': 1,
 'bbcsearch': 1,
 'because': 2,
 'before': 2,
 'beijing': 1,
 'blogger': 1,
 'bloomberg': 5,
 "bloomberg's": 4,
 'bloombergthe': 1,
 'boards': 1,
 'body': 1,
 'bolsonaro': 1,
 'bought': 1,
 'brazil': 1,
 'breach': 1,
 'british': 1,
 'business': 5,
 "buy'": 1,
 'called': 2,
 'calling': 1,
 'campaign': 1,
 'canada': 1,
 'candidate': 1,
 'caption': 2,
 'carried': 1,
 'carry': 1,
 'carrying': 1,
 'caught': 1,
 'centres': 1,
 'chain': 3,
 'chains': 1,
 'charts': 1,
 'check': 2,
 'chef': 1,
 'chief': 1,
 'china': 3,
 'chinese': 2,
 'chips': 3,
 'chocolate': 1,
 'chop': 1,
 'cia': 1,
 'circuit': 1,
 'claims': 5,
 'claimsin': 1,
 'code': 1,
 'common': 1,
 'companies': 6,
 'company': 2,
 'companyamazon': 1,
 'components': 1,
 'compromised': 4,
 'computer': 4,
 'computers': 1,
 'concernbloomberg': 1,
 'conducted': 1,
 'confessions': 1,
 'confirms': 1,
 'connected': 2,
 'consistently': 1,
 'consulate': 1,
 'contact': 1,
 'contacted': 1,
 'content': 1,
 'contentaccessibility': 1,
 'copy': 1,
 'copyright': 2,
 'countered': 1,
 'country': 1,
 'current': 1,
 'customer': 1,
 'cyber-attacks': 4,
 'data': 5,
 "daughter'": 1,
 'deep': 1,
 'defense': 1,
 'denials': 1,
 'denied': 1,
 'department': 1,
 'design': 1,
 'desired': 1,
 'detailed': 1,
 'detained': 1,
 'developing': 1,
 'devices': 1,
 "digitaltasterlocaltomorrow's": 2,
 'disabled': 1,
 'discovery': 1,
 'doctored': 1,
 'draws': 1,
 'drones': 1,
 'during': 2,
 'each': 1,
 'east': 1,
 'economy': 1,
 'education': 1,
 'elaborate': 1,
 'elemental': 1,
 'elsewhere': 1,
 'emerged': 1,
 'ending': 1,
 'england': 1,
 'ensuring': 1,
 'entertainment': 2,
 'entrepreneurship': 1,
 "escobar's": 1,
 'europe': 1,
 'evidence': 4,
 'explainers': 2,
 'explore': 1,
 'external': 4,
 'extreme': 1,
 'facebook': 5,
 'factory': 1,
 'factual': 1,
 'far-right': 1,
 'fears': 1,
 'features': 1,
 'federal': 1,
 'feeling': 2,
 'fighting': 1,
 'food': 1,
 'foreign': 1,
 'from': 2,
 'full': 2,
 'gathered': 1,
 'gave': 1,
 'get': 2,
 'getting': 1,
 'getty': 1,
 'give': 1,
 'global': 3,
 'go': 1,
 'government': 1,
 'gratuitous': 1,
 'guidancecontact': 1,
 'hackers': 1,
 "hackerschina's": 1,
 'handling': 1,
 'happens': 1,
 'harbouring': 1,
 'hardware': 2,
 'health': 2,
 'heat?': 1,
 'helpbbc': 1,
 'helpparental': 1,
 'hits': 2,
 'homepageaccessibility': 1,
 'https//wwwbbccom/news/technology-45747983stm': 1,
 "i've": 1,
 'idnotificationshomenewssportweatheriplayertvradiocbbccbeebiesfoodbitesiz
emusicearthartsmake': 1,
 'image': 4,
 'images': 1,
 'in?': 1,
 'incidenteach': 1,
 'including': 1,
 'information': 1,
 'inquiries': 1,
 'inserted': 1,
 'insiders': 1,
 'intelligence': 1,
 'internal': 1,
 'interpol': 1,
 'investigation': 3,
 'investigations': 1,
 'involved': 1,
 'ireland': 1,
 'issue': 2,
 'jordan': 1,
 'kicked': 1,
 'kind': 1,
 'latin': 1,
 'lead': 1,
 'led': 1,
 'lengthy': 1,
 'life': 1,
 'linkedin': 3,
 'linking': 1,
 'links': 2,
 'linksskip': 1,
 'location': 1,
 'logistics': 1,
 'long-running': 1,
 'looks': 1,
 'lovechild': 1,
 'lyrics': 2,
 'major': 1,
 'malicious': 1,
 'maliciously': 1,
 'manipulating': 1,
 'manufactured': 1,
 'manufacturing': 1,
 'market': 1,
 'meets': 1,
 'messenger': 8,
 'michael': 1,
 'michelin': 1,
 'micro': 6,
 'middle': 1,
 'military': 1,
 'ministry': 1,
 'mobile': 1,
 'modificationsapple': 1,
 'modified': 1,
 "moon'": 1,
 'multiple': 1,
 'murder': 1,
 'my': 2,
 'mystery': 1,
 'n': 1,
 'national': 1,
 'navigation': 2,
 'never': 1,
 'newsbeat': 2,
 'newsletterscopyright': 1,
 'october': 4,
 'offered': 1,
 'officials': 1,
 'onboard': 1,
 'once': 1,
 'onions?': 1,
 'open': 2,
 'our': 1,
 'pablo': 1,
 'panel': 1,
 'particular': 1,
 'pcs': 1,
 'pentagon': 1,
 'personalised': 1,
 'pictures': 2,
 'pinterest': 3,
 'placed': 1,
 'plants': 1,
 'policycookiesaccessibility': 1,
 'politics': 1,
 'polls': 2,
 'pop': 1,
 'president': 1,
 'presidential': 1,
 'probe': 2,
 "product's": 1,
 'products': 1,
 'prostitute': 1,
 'quiz': 2,
 'reaches': 1,
 'read': 2,
 'reality': 2,
 'record': 1,
 'refuting': 1,
 'rejected': 1,
 'related': 1,
 'relating': 1,
 'relationships': 1,
 'relaxes': 2,
 'released': 1,
 'removing': 1,
 'repeatedly': 1,
 'reported': 1,
 'reporters': 3,
 'reports': 2,
 'response': 1,
 'responses': 1,
 'responsible': 1,
 'reykjavik': 1,
 'right': 1,
 'rigorous': 1,
 'riley': 1,
 'robertson': 1,
 'round': 1,
 'rules': 2,
 'running': 1,
 'safety': 1,
 'saidapple': 1,
 'saudi': 1,
 'saying': 2,
 'science': 2,
 'scotland': 1,
 'sea': 1,
 'second': 1,
 'sections': 2,
 'security': 3,
 'seen': 1,
 'selected': 2,
 'september': 1,
 'server': 1,
 'servers': 4,
 'services': 1,
 'set': 1,
 'share': 20,
 'sharing': 2,
 'siphoned': 1,
 'sites': 1,
 'six': 1,
 'software': 2,
 'some': 1,
 'sometimes': 2,
 'songs': 1,
 'special': 2,
 'spies': 1,
 'spy': 2,
 'spying': 1,
 'started': 1,
 'statement': 2,
 'stolen': 1,
 'stopped': 1,
 'stories': 3,
 'story': 4,
 'strong': 1,
 'substance': 1,
 'suggest': 1,
 'super': 5,
 'supply': 3,
 'support': 3,
 'task': 1,
 'tasting': 1,
 'tech': 2,
 'technology': 3,
 'testimony': 1,
 'testing': 1,
 'themit': 1,
 'then': 1,
 'through': 2,
 'throws': 1,
 'ticket': 1,
 'time': 2,
 'times': 1,
 'tiny': 1,
 'took': 1,
 'top-secret': 1,
 'topicschinacyber-attackamazonapple': 1,
 'town': 1,
 'trade': 1,
 'travel': 1,
 'trump': 2,
 'trump’s': 1,
 'trust': 1,
 'tv': 3,
 'twitter': 5,
 'uk': 3,
 'uncovered': 1,
 'understanding': 1,
 'unite': 1,
 'untruein': 1,
 'up': 2,
 'useabout': 1,
 'using': 3,
 'vague': 1,
 'via': 1,
 'video': 4,
 'virtually': 1,
 'wales': 1,
 'warns': 3,
 'warships': 2,
 'we': 3,
 "we've": 1,
 'well': 2,
 'were': 4,
 'whatsapp': 4,
 'wide': 1,
 'wide-ranging': 1,
 'window': 2,
 'wine': 1,
 'world': 6,
 "world's": 1,
 'worldmenusearchsearch': 1,
 'worldterms': 1,
 'writer': 1,
 'wrong?': 1,
 'year-long': 1,
 "you've": 1}
```

In [ ]:

In [25]:

```
# 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: 50

OK. The 50 most common words are as follows

share :  20
bbc :  11
messenger :  8
apple :  7
world :  6
amazon :  6
companies :  6
micro :  6
attack :  5
business :  5
facebook :  5
twitter :  5
bloomberg :  5
data :  5
claims :  5
super :  5
video :  4
october :  4
whatsapp :  4
external :  4
image :  4
were :  4
compromised :  4
computer :  4
servers :  4
bloomberg's :  4
evidence :  4
cyber-attacks :  4
story :  4
uk :  3
stories :  3
& :  3
tv :  3
reporters :  3
technology :  3
pinterest :  3
linkedin :  3
agencies :  3
chips :  3
article :  3
support :  3
investigation :  3
warns :  3
supply :  3
chain :  3
security :  3
using :  3
china :  3
global :  3
we :  3
```

In [26]:

```
# 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[26]:

	Word	Count
0	share	20
1	bbc	11
2	messenger	8
3	apple	7
4	world	6
5	amazon	6
6	companies	6
7	micro	6
8	attack	5
9	business	5

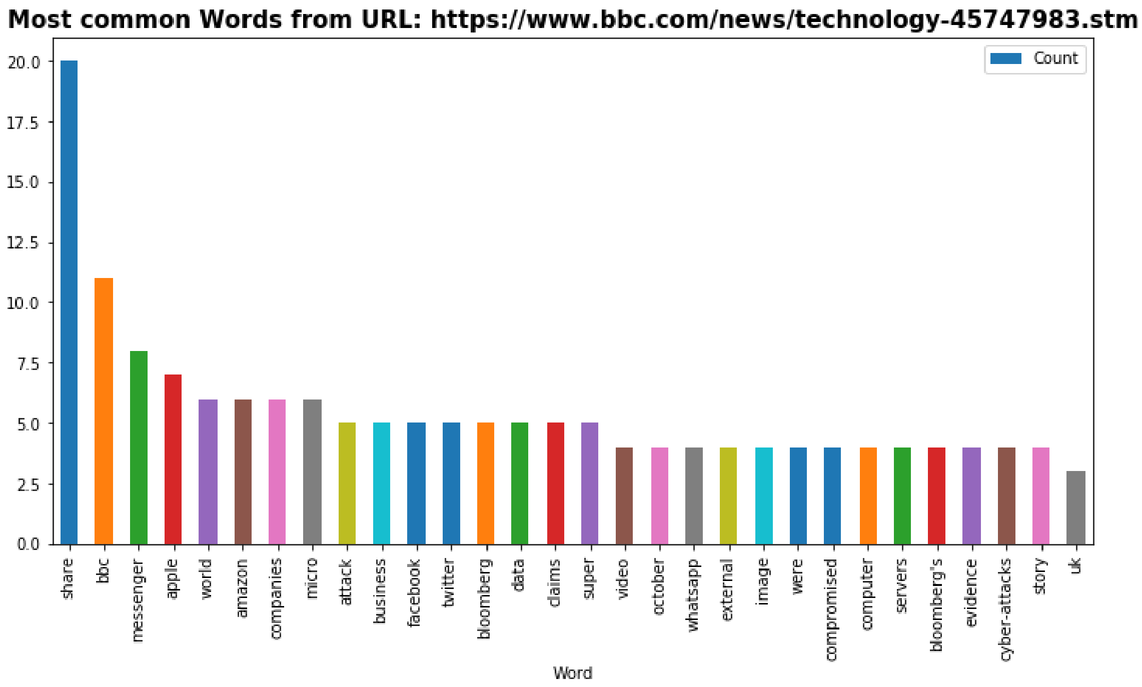
In [27]:

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

In [28]:

```
# 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 [29]:

```
wrds = df.Word

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

Out[29]:

```
0        share
1          bbc
2    messenger
3        apple
4        world
Name: Word, dtype: object
```

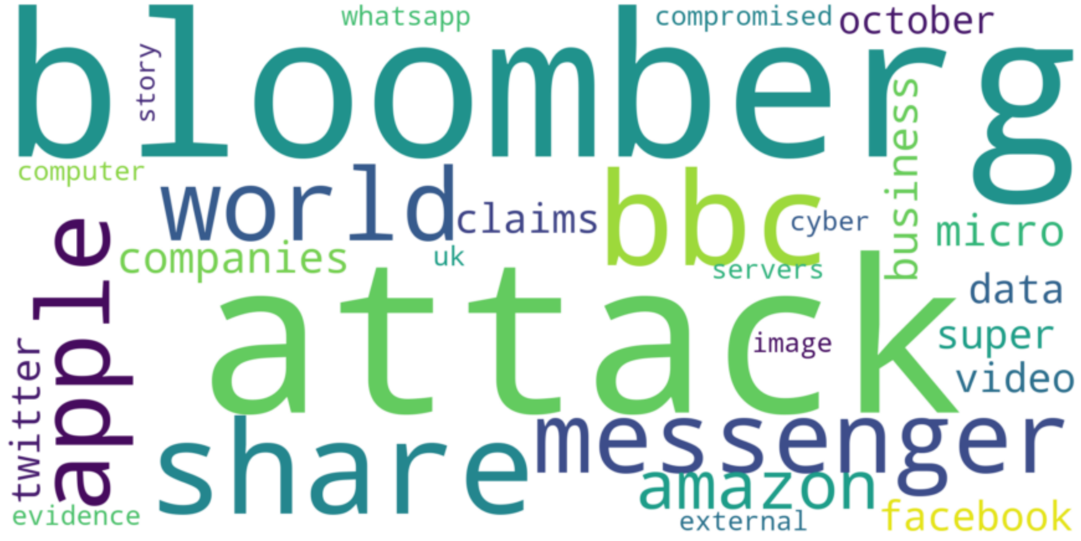
In [30]:

```
# 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'});
```

Most common Words from URL: https://www.bbc.com/news/technology-45747983.stm



In [ ]: