Webscraping_NLP

January 6, 2021

0.1 Task:

0.1.1 Fetch data from php.net website, clean it and give frequency of each word.

[58]:
```
#Importing required python libraries
from bs4 import BeautifulSoup
import urllib.request
import nltk
import re
import matplotlib.pyplot as plt
from wordcloud import WordCloud
```

[9]:
```
res=urllib.request.urlopen('http://php.net/')
```

[10]:
```
html=res.read()
soup=BeautifulSoup(html,'html.parser')
```

[28]:
```
text=soup.get_text()
```

[31]:
```
#Replacing all new line char \n with space
text=text.replace('\n',' ').strip()
```

[36]:
```
#Replacing multiple spaces with single space
text=re.sub(r'\s+',' ',text)
```

[38]:
```
#Extracting only characters from the text and removing digits and other␣
↪punctuations.
text=re.sub(r'[^a-zA-Z ]','',text)
```

[42]:
```
#Splitting the sentences to list of words and converting all words to lowercase
text_list=[x.lower() for x in text.split()]
```

[44]:
```
stop_words=nltk.corpus.stopwords.words('english')
```

[54]:
```
#Removing all words with length less than or equal to 2 or are stopwords
words=[]
for word in text_list:
    if(len(word)<=2 or word in stop_words):
```

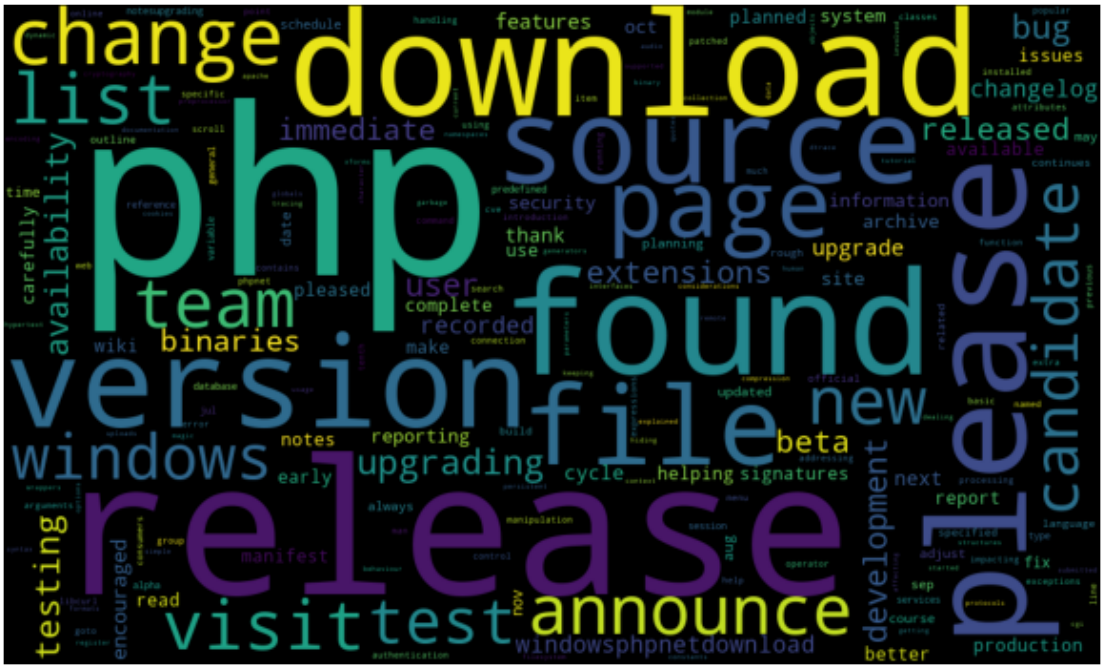
```
        continue
    else:
        words.append(word)
```

[56]:
```
extracted_words=' '.join(words)
```

[68]:
```
wordcloud=WordCloud(width = 1000, height = 600,collocations=False).
↪generate(extracted_words)
```

[70]:
```
# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```



[73]:
```
freq = nltk.FreqDist(words)
```

[78]:
```
print(freq.items())
```

```
dict_items([('php', 158), ('hypertext', 1), ('preprocessor', 1), ('downloads',
41), ('documentation', 1), ('get', 1), ('involved', 1), ('help', 2), ('getting',
1), ('started', 1), ('introduction', 2), ('simple', 1), ('tutorial', 1),
('language', 4), ('reference', 2), ('basic', 2), ('syntax', 1), ('types', 2),
```

('variables', 2), ('constants', 1), ('expressions', 2), ('operators', 1),
('control', 2), ('structures', 1), ('functions', 1), ('classes', 2), ('objects',
1), ('namespaces', 1), ('errors', 1), ('exceptions', 2), ('generators', 1),
('attributes', 2), ('references', 1), ('explained', 1), ('predefined', 3),
('interfaces', 1), ('context', 1), ('options', 1), ('parameters', 1),
('supported', 1), ('protocols', 1), ('wrappers', 1), ('security', 11),
('general', 3), ('considerations', 1), ('installed', 2), ('cgi', 1), ('binary',
1), ('apache', 1), ('module', 1), ('session', 2), ('filesystem', 1),
('database', 3), ('error', 1), ('reporting', 11), ('using', 2), ('register', 1),
('globals', 1), ('user', 2), ('submitted', 1), ('data', 1), ('magic', 1),
('quotes', 1), ('hiding', 1), ('keeping', 1), ('current', 1), ('features', 13),
('http', 1), ('authentication', 2), ('cookies', 1), ('sessions', 1), ('dealing',
1), ('xforms', 1), ('handling', 2), ('file', 22), ('uploads', 1), ('remote', 1),
('files', 11), ('connection', 1), ('persistent', 1), ('connections', 1),
('command', 2), ('line', 2), ('usage', 1), ('garbage', 1), ('collection', 1),
('dtrace', 1), ('dynamic', 1), ('tracing', 1), ('function', 1), ('affecting',
1), ('phps', 1), ('behaviour', 1), ('audio', 1), ('formats', 1),
('manipulation', 2), ('services', 3), ('specific', 2), ('extensions', 16),
('compression', 1), ('archive', 11), ('cryptography', 1), ('date', 7), ('time',
7), ('related', 4), ('system', 11), ('human', 1), ('character', 1), ('encoding',
1), ('support', 1), ('image', 1), ('processing', 2), ('generation', 1), ('mail',
1), ('mathematical', 1), ('nontext', 1), ('mime', 1), ('output', 1), ('process',
1), ('search', 2), ('engine', 1), ('server', 1), ('text', 1), ('variable', 1),
('type', 1), ('web', 2), ('windows', 20), ('xml', 1), ('gui', 1), ('keyboard',
1), ('shortcuts', 1), ('next', 12), ('menu', 2), ('item', 2), ('previous', 2),
('man', 2), ('page', 33), ('scroll', 2), ('bottom', 1), ('top', 1), ('goto', 2),
('homepage', 1), ('searchcurrent', 1), ('focus', 1), ('box', 1), ('popular', 2),
('generalpurpose', 1), ('scripting', 1), ('especially', 1), ('suited', 1),
('development', 17), ('fast', 1), ('flexible', 1), ('pragmatic', 1), ('powers',
1), ('everything', 1), ('blog', 1), ('websites', 1), ('world', 1), ('download',
11), ('release', 97), ('notesupgrading', 3), ('nov', 8), ('released', 16),
('team', 25), ('announces', 15), ('immediate', 15), ('availability', 17),
('marks', 1), ('latest', 1), ('major', 1), ('comes', 1), ('numerous', 1),
('improvements', 1), ('new', 12), ('union', 1), ('named', 2), ('arguments', 2),
('match', 1), ('constructor', 1), ('property', 1), ('promotion', 1),
('nullsafe', 1), ('operator', 1), ('weak', 1), ('maps', 1), ('compilation', 1),
('much', 2), ('take', 1), ('look', 1), ('announcement', 1), ('addendum', 1),
('information', 11), ('source', 40), ('please', 46), ('visit', 25), ('binaries',
15), ('found', 51), ('windowsphpnetdownload', 15), ('list', 26), ('changes',
26), ('recorded', 15), ('changelog', 15), ('migration', 1), ('guide', 1),
('available', 10), ('manual', 1), ('consult', 1), ('detailed', 1), ('backward',
1), ('incompatible', 1), ('many', 1), ('thanks', 2), ('contributors', 1),
('supporters', 1), ('bug', 18), ('fix', 8), ('users', 15), ('encouraged', 13),
('upgrade', 14), ('version', 45), ('candidate', 20), ('testing', 19),
('pleased', 10), ('announce', 10), ('eleventh', 1), ('extra', 2), ('unplanned',
1), ('planning', 6), ('adjust', 6), ('however', 7), ('may', 6), ('change', 6),
('course', 6), ('cycle', 12), ('updated', 6), ('schedule', 6), ('always', 6),
('wiki', 10), ('carefully', 10), ('test', 20), ('report', 10), ('issues', 10),

('use', 10), ('production', 10), ('early', 10), ('read', 10), ('news', 11), ('upgrading', 20), ('complete', 10), ('notes', 10), ('also', 10), ('planned', 10), ('signatures', 10), ('manifest', 10), ('site', 10), ('thank', 10), ('helping', 10), ('make', 10), ('better', 10), ('tenth', 2), ('oct', 11), ('ninth', 1), ('eighth', 1), ('sep', 6), ('beta', 16), ('seventh', 1), ('point', 2), ('would', 1), ('normally', 1), ('still', 1), ('finalizing', 1), ('jit', 1), ('squaring', 1), ('away', 1), ('weve', 1), ('opted', 1), ('plans', 1), ('start', 1), ('two', 1), ('weeks', 1), ('sixth', 1), ('continues', 4), ('rough', 4), ('outline', 4), ('specified', 4), ('aug', 7), ('fifth', 1), ('fourth', 1), ('jul', 3), ('alpha', 3), ('third', 1), ('impacting', 2), ('official', 4), ('builds', 2), ('running', 2), ('build', 2), ('contains', 2), ('patched', 2), ('libcurl', 2), ('addressing', 2), ('cve', 2), ('consumers', 2), ('functionally', 1), ('identical', 1), ('necessary', 1), ('older', 1), ('entries', 1), ('upcoming', 1), ('conferenceslaracon', 1), ('online', 2), ('conference', 1), ('group', 2), ('events', 1), ('special', 1), ('social', 1), ('media', 1), ('officialphp', 1), ('copyright', 1), ('phpnet', 2), ('contact', 1), ('sites', 1), ('privacy', 1), ('policy', 1)])