Forward School

Program Code: J620-002-4:2020

Program Name: FRONT-END SOFTWARE DEVELOPMENT

Title : Text Visualization with Wordcloud

Name: Chuay Xiang Ze

IC Number: 021224070255

Date : 6/7/2023

Introduction : Learning how to visualize text using Wordcloud

Conclusion : Managed to display dictionary of texts based on text files and also texts in shape of png images.

P17 - Visualizing Text with Word Cloud

Word Cloud

What is a word cloud?

Data visualizations (like charts, graphs, infographics, and more) one of the many ways to communicate important information at a glance, but what if the raw data is text-based?

Word clouds (also known as text clouds or tag clouds): the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud.

A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is.

Also known as tag clouds or text clouds, these are ideal ways to pull out the most pertinent parts of textual data, from blog posts to databases. They can also help business users compare and contrast two different pieces of text to find the wording similarities between the two.

Useful for quick summary of common customer feedback, text documents, identifying new SEO terms to target.

https://pypi.org/project/wordcloud/ (https://pypi.org/project/wordcloud/)

Know how to search for packages?

https://en.wikipedia.org/wiki/Tag_cloud (https://en.wikipedia.org/wiki/Tag_cloud)

References:

https://amueller.github.io/word_cloud/ (https://amueller.github.io/word_cloud/)

https://github.com/amueller/word_cloud (https://github.com/amueller/word_cloud)

https://www.kaggle.com/agisga/word-clouds (https://www.kaggle.com/agisga/word-clouds)

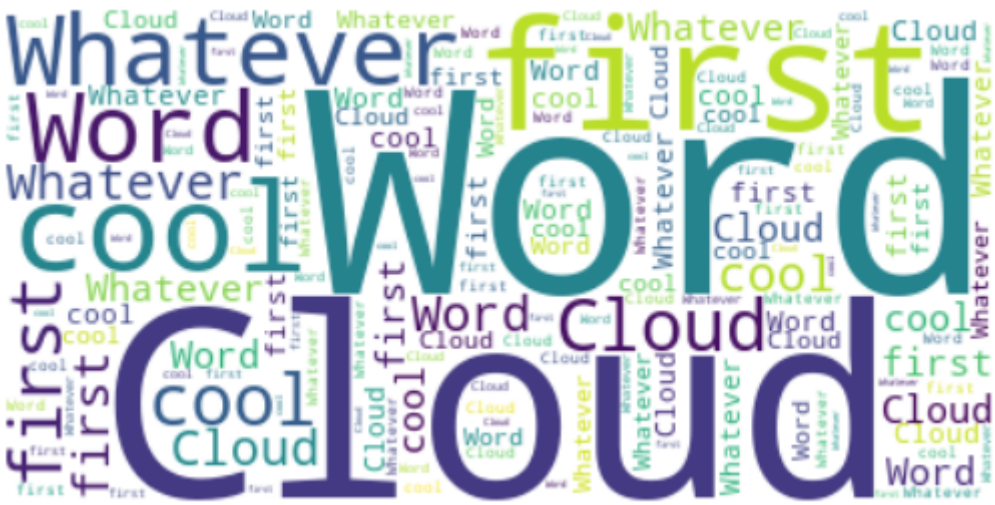
https://www.wordclouds.com/ (https://www.wordclouds.com/)

Installation

conda install -c conda-forge wordcloud

In [1]:

```
import matplotlib.pyplot as plt
from wordcloud import WordCloud

text = "This is my first Word Cloud, Word Cloud is cool. Whatever this is"

wc = WordCloud()
wc = WordCloud(background_color="white", repeat=True)

wc.generate(text)

plt.axis("off")
plt.imshow(wc, interpolation="bilinear")
plt.show()
```



In [2]:

```
from wordcloud import WordCloud, STOPWORDS

STOPWORDS
```

Out[2]:

```
{'a',
 'about',
 'above',
 'after',
 'again',
 'against',
 'all',
 'also',
 'am',
 'an',
 'and',
 'any',
 'are',
 "aren't",
 'as',
 'at',
 'be',
 'because',
```

Let's get real world data

From Wikipedia

conda install -c conda-forge wikipedia

In [ ]:

```
import sys
import wikipedia
from wordcloud import WordCloud, STOPWORDS

inputstring = str(input('Enter the title; '))

title = wikipedia.search(inputstring)[0]

page = wikipedia.page(title)

text = page.content
```

In [ ]:

```
print(text)
```

In [ ]:

```
import matplotlib.pyplot as plt

wordcloud  = WordCloud(background_color='black', max_words=200, stopwords=STOPWORDS)

wordcloud.generate(text)

plt.imshow(wordcloud,interpolation='bilinear')
plt.axis("off")
plt.show()
```

2. From PDF File

In [2]:

```
import requests

url = 'https://www.agc.gov.my/agcportal/common//uploads/publication/391/2020_11_13_DKICT

# Download the PDF
myfile = requests.get(url, allow_redirects=True, verify=False)

open('./IT_Security_Policy_for_AGC.pdf', 'wb').write(myfile.content)
```

```
C:\Anaconda\envs\python-dscourse\lib\site-packages\urllib3\connectionpool.
py:1056: InsecureRequestWarning: Unverified HTTPS request is being made to
host 'www.agc.gov.my'. Adding certificate verification is strongly advise
d. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-w
arnings (https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-
warnings)
  warnings.warn(
```

Out[2]:

```
1485266
```

In [3]:

```
# Convert PDF to Text
import PyPDF2

with open('IT_Security_Policy_for_AGC.pdf','rb') as pdf_file, open('IT_Security_Policy_f
    read_pdf = PyPDF2.PdfFileReader(pdf_file)
    number_of_pages = read_pdf.getNumPages()
    for page_number in range(number_of_pages):
        page = read_pdf.getPage(page_number)
        page_content = page.extractText()
        text_file.write(page_content)
```

In [4]:

```
page_content
```

Out[4]:

```
'DASAR KESELAMATAN TE KNOLOGI MAKLUMAT JAB ATAN PEGUAM NEGARA   \nTARIKH :
15 FEBRUARI  2018 MUKA SURAT  74 DARI 75 \n o) Akta Rahsia Rasmi 1972;  \n
\np) Akta Jenayah Komputer 1997;  \n \nq) Akta Hak Cipta (Pindaan) Tahun 1
997;  \n \nr) Akta Komunikasi dan Multimedia 1998;  \n \ns) Perintah -Peri
ntah Am;  \n \nt)  Arahan Perbendaharaan;  \n \nu) Arahan Teknologi Maklum
at 2007;  \n \nv)  Garis Panduan Keselamatan AGC 2004;  \n \nw) Standard O
perating Procedure (SOP) ICT AGC;  \n \nx) Surat Pekeliling Am Bilangan 3
Tahun 2009 – Garis Panduan Penilaian Tahap \nKeselamatan Rangkaian dan Sis
tem ICT Sektor Awam yang bertarikh 17 \nNovember 2009;  \n \ny) Surat Arah
an Peguam Negara AGC – Pengurusan Kesinambungan \nPerkhidmatan Ag ensi Sek
tor Awam yang bertarikh 22 Januari 2010.  \n '
```

Alternative PDF libraries

https://anaconda.org/anaconda/repo (https://anaconda.org/anaconda/repo)

http://mstamy2.github.io/PyPDF2/ (http://mstamy2.github.io/PyPDF2/)

https://pypi.org/project/pdftotext/ (https://pypi.org/project/pdftotext/)

https://realpython.com/pdf-python/ (https://realpython.com/pdf-python/)

Downloading Files

https://dzone.com/articles/simple-examples-of-downloading-files-using-python
(https://dzone.com/articles/simple-examples-of-downloading-files-using-python)

In [ ]:

```
# %matplotlib inline
```

In [5]:

```
from wordcloud import WordCloud, STOPWORDS

# Read the whole text.
text = open('./IT_Security_Policy_for_AGC.txt', encoding='utf-8').read()
text = text.replace ('  ',' ')

# Generate a word cloud image
wordcloud = WordCloud().generate(text)

# Display the generated image:

# the matplotlib way:
import matplotlib.pyplot as plt
plt.axis("off")
plt.imshow(wordcloud, interpolation='bilinear')
plt.show()

# The pil way (if you don't have matplotlib)
# from IPython.display import Image
# pil_img = wordcloud.to_image()
# display(pil_img)
```

In [8]:

```
# Generate a word cloud image
wordcloud = WordCloud().generate(text)

# Display the generated image:
plt.figure(figsize=(10,10)) #inches
plt.axis("off")
plt.imshow(wordcloud, interpolation='bilinear')

plt.show()

# note image size generated and the canvas size of plot
# https://matplotlib.org/3.2.1/api/_as_gen/matplotlib.pyplot.figure.html
```

In [9]:

```
# Generate a word cloud image
wordcloud = WordCloud(width=800, height=400).generate(text)
#wordcloud = WordCloud(width=3600, height=1600).generate(text)

# Display the generated image:
plt.figure(figsize=(10,10)) # inches
plt.axis("off")
plt.imshow(wordcloud, interpolation='bilinear')
plt.show()

# note image size generated and the canvas size of plot
```

In [10]:

```
# lower max_font_size
wordcloud = WordCloud(max_font_size=20).generate(text)

# Display the generated image:
plt.figure()
plt.axis("off")
plt.imshow(wordcloud, interpolation="bilinear")
plt.show
```

Out[10]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

In [11]:

```
# Change font size, Background Color
wordcloud = WordCloud(max_font_size=50, background_color='white').generate(text)
plt.figure()
plt.axis("off")
plt.imshow(wordcloud, interpolation="bilinear")
plt.show
```

Out[11]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

In [12]:

```
# lower font size, maximum words, Background Color
wordcloud = WordCloud(max_font_size=50, max_words=10,background_color='white').generate(
plt.figure()
plt.axis("off")
plt.imshow(wordcloud, interpolation="bilinear")
plt.show
```

Out[12]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

In [13]:

```
from wordcloud import STOPWORDS

# Create stopword list:
stopwords = set(STOPWORDS)

wordcloud = WordCloud(stopwords=stopwords, max_font_size=50, max_words=10,background_col
plt.figure()
plt.axis("off")
plt.imshow(wordcloud, interpolation="bilinear")
plt.show
```

Out[13]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

In [14]:

```
from wordcloud import STOPWORDS

# Create stopword list:
stopwords = set(STOPWORDS)
stopwords.update(["yang","di","sabah","sarawak", "section","force","clause", "act","pert
# stop_words = list(stopwords)+["yang", "di","sabah sarawak", "section force", "force cl

wordcloud = WordCloud(stopwords=stopwords, max_font_size=50, max_words=10,background_col
plt.figure()
plt.axis("off")
plt.imshow(wordcloud, interpolation="bilinear")
plt.show
```

Out[14]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

In [15]:

```
stopwords
```

Out[15]:

```
{'a',
 'about',
 'above',
 'act',
 'after',
 'again',
 'against',
 'agong',
 'all',
 'also',
 'am',
 'an',
 'and',
 'any',
 'are',
 "aren't",
 'as',
 'at',
```

In [16]:

```
from wordcloud import WordCloud, STOPWORDS

testtext = 'yang di is'

# Create stopword list:
stopwords = STOPWORDS
stop_words = ['yang'] + list(stopwords)

wordcloud = WordCloud(stopwords=stop_words, max_font_size=50, max_words=10,background_co

plt.figure()
plt.axis("off")
plt.imshow(wordcloud, interpolation="bilinear")
plt.show
```

Out[16]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

Mask

Change the layout

Generate a Numpy grid

In [17]:

```
# Add Mask
import numpy as np

x, y = np.ogrid[:300, :300]

mask = (x - 150) ** 2 + (y - 150) ** 2 > 130 ** 2
mask = 255 * mask.astype(int)

wordcloud = WordCloud(max_font_size=50, max_words=10,background_color='white', mask=mask
plt.figure()
plt.axis("off")
plt.imshow(wordcloud, interpolation="bilinear")
plt.show
plt.savefig('mask.png')
```

Mask from another Image

First find or create an Image

Eg.

1. Use Paint and save it as mask.png

In [18]:

```
from IPython.display import display, Image
display(Image(filename='./mask.png'))
```

In [17]:

```
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import numpy as np
from PIL import Image
import matplotlib.pyplot as plt

mask = np.array(Image.open('./mcfc.png'))

color= ImageColorGenerator(mask)

wordcloud = WordCloud(width=50,
                      height=50,
                      max_words=50000,
                      mask=mask,
                      stopwords=STOPWORDS,
                      background_color='white',
                      random_state=42).generate(text)

plt.figure(figsize=(20,20)) # inches
plt.axis("off")
plt.imshow(wordcloud.recolor(color_func=color), interpolation='bilinear')
plt.show()
```

Read up

https://matplotlib.org/gallery/images_contours_and_fields/interpolation_methods.html
(https://matplotlib.org/gallery/images_contours_and_fields/interpolation_methods.html)

In [ ]:

```
wordcloud.to_file('')
```

2. Or download an Image

Flag of Malaysia

User Google Search

Find Images with larger sizes

Eg. https://en.wikipedia.org/wiki/Flag_of_Malaysia#/media/File:Flag_of_Malaysia.svg
(https://en.wikipedia.org/wiki/Flag_of_Malaysia#/media/File:Flag_of_Malaysia.svg)

In [20]:

```
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import numpy as np
from PIL import Image
import matplotlib.pyplot as plt

mask = np.array(Image.open('./1920px-Flag_of_Malaysia.svg.png'))

color= ImageColorGenerator(mask)

wordcloud = WordCloud(width=1920,
                      height=1080,
                      max_words=400,
                      mask=mask,
                      stopwords=STOPWORDS,
                      background_color='white',
                      random_state=42).generate(text)

plt.figure(figsize=(10,10)) # inches
plt.axis("off")
plt.imshow(wordcloud.recolor(color_func=color),interpolation='bilinear')
plt.show()
```

In [21]:

```
# Save to File
wordcloud.to_file('MalaysiaWordCloud.png')
```

Out[21]:

```
<wordcloud.wordcloud.WordCloud at 0x1b1e31fd190>
```

Try all the examples below

Python script to search google and produce a word cloud from the abstracts of the first page of results

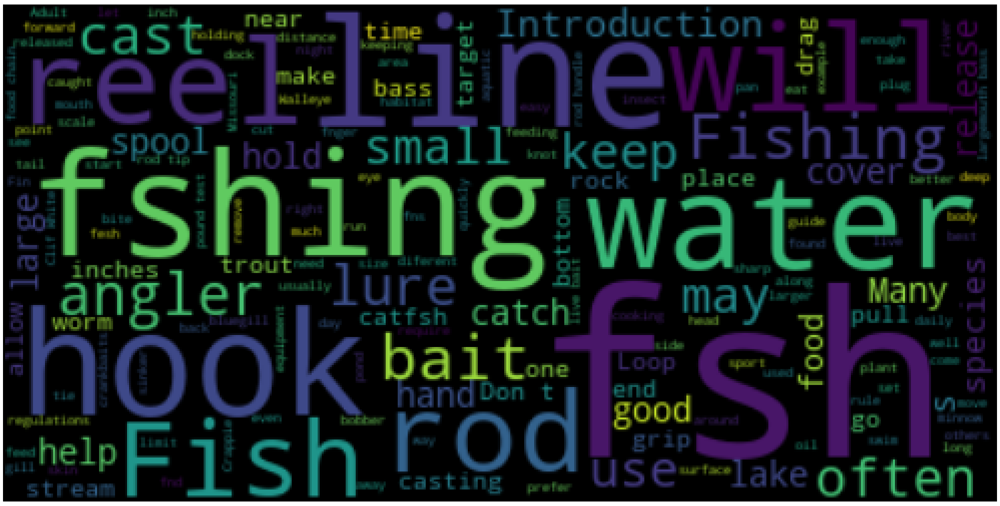
https://github.com/charlie9578/googleWordCloud (https://github.com/charlie9578/googleWordCloud)

In [22]:

```
from wordcloud import WordCloud, STOPWORDS
from PIL import Image
import urllib
import requests
import numpy as np
import matplotlib.pyplot as plt

words = 'access guest guest apartment area area bathroom bed bed bed bed bed bedroom blo
mask = np.array(Image.open(requests.get('http://www.clker.com/cliparts/O/i/x/Y/q/P/yello

# This function takes in your text and your mask and generates a wordcloud.
def generate_wordcloud(words, mask):
    word_cloud = WordCloud(width = 512, height = 512, background_color='white', stopword
    plt.figure(figsize=(10,8),facecolor = 'white', edgecolor='blue')
    plt.imshow(word_cloud)
    plt.axis('off')
    plt.tight_layout(pad=0)
    plt.show()

#Run the following to generate your wordcloud
generate_wordcloud(words, mask)
```

In [36]:

```
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import matplotlib.pyplot as plt
import numpy as np
from PIL import Image
# Read the whole text.
text = open('./An Introduction to Fishing.txt', encoding='utf-8').read()
text = text.replace ('  ',' ')

# Generate a word cloud image
wordcloud = WordCloud().generate(text)

# Display the generated image:

# the matplotlib way:

plt.axis("off")
plt.imshow(wordcloud, interpolation='bilinear')
plt.show()

wordcloud = WordCloud(background_color='white').generate(text)
plt.figure()
plt.axis("off")
plt.imshow(wordcloud, interpolation="bilinear")
plt.show

# The pil way (if you don't have matplotlib)
# from IPython.display import Image
# pil_img = wordcloud.to_image()
# display(pil_img)

mask = np.array(Image.open('./vaporeon.png'))

color= ImageColorGenerator(mask)

wordcloud = WordCloud(width=50,
                      height=50,
                      max_font_size=100,
                      max_words=5000000,
                      mask=mask,
                      stopwords=STOPWORDS,
                      background_color='white',
                      random_state=42).generate(text)

plt.figure(figsize=(20,20)) # inches
plt.axis("off")
plt.imshow(wordcloud.recolor(color_func=color), interpolation='bilinear')
plt.show()
```


reel
line
will
fshing
water
fsh
hook
Fish
rod
Fishing
Introduction
cast
small
keep
spool
hold
angler
lure
may
Many
bait
catch
good
often
use
lake
help
hand
release
species
cover
place
large
inches
trout
bottom
food
worm
allow
stream
casting
catfsh
pull
end
Loop
grip
go
bass
make
near
time
target
drag
rock
one
Don t


water
fsh
reel
Fish
hook
line
rod
fshing
will
Fishing
angler
spool
release
use
lake
large
catch
help
place
often
good
cover
lure
small
bait
Loop
food
target
Don t
trout
stream
may
Introduction
hand
Many
hold
end
cast
bass
make
near
casting
pull
keep
worm
species
inches
catfsh
rock
allow
one
drag
time
bottom
go
grip
regulations
bobber


```
# url: https://www.deviantart.com/proteusiii/art/Vaporeon-Sitting-PNG-351031719
```

_cloud/)

pdf

.pdf)