Forward School

Program Code: J620-002-4:2020

Program Name: FRONT-END SOFTWARE DEVELOPMENT

Title : Text Visualization with Wordcloud

Name: Ooi Caaron

IC Number: 990701-07-5837

Date : 4/7/23

Introduction : Learning how to create a word cloud using python programming

Conclusion : Still need to practice more and do revision

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)

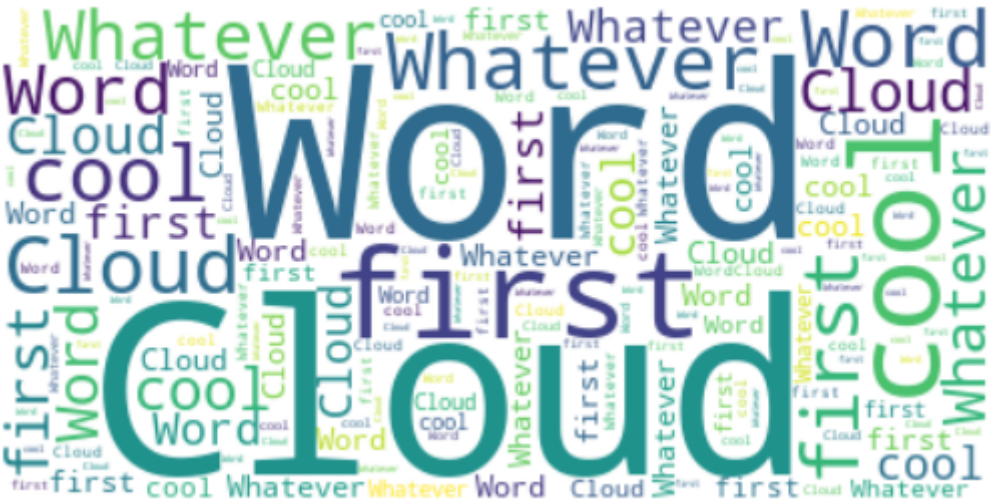
https://www.wordclouds.com/ (https://www.wordclouds.com/)

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 [39]:

```
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
```

```
Enter the title; Arsenal
```

In [40]:

```
print(text)
```

```
Arsenal Football Club is an English professional football club based in
Islington, London. Arsenal play in the Premier League, the top flight o
f English football. The club has won 13 league titles (including one un
beaten title), a record 14 FA Cups, two League Cups, 16 FA Community Sh
ields, one European Cup Winners' Cup, and one Inter-Cities Fairs Cup. I
n terms of trophies won, it is the third-most successful club in Englis
h football.
Arsenal was the first club from the South of England to join the Footba
ll League in 1893, and they reached the First Division in 1904. Relegat
ed only once, in 1913, they continue the longest streak in the top divi
sion, and have won the second-most top-flight matches in English footba
ll history. In the 1930s, Arsenal won five League Championships and two
FA Cups, and another FA Cup and two Championships after the war. In 197
0–71, they won their first League and FA Cup Double. Between 1989 and 2
005, they won five League titles and five FA Cups, including two more D
oubles. They completed the 20th century with the highest average league
position. Between 1998 and 2017, Arsenal qualified for the UEFA Champio
ns League for nineteen consecutive seasons.
Herbert Chapman, who changed the fortunes of Arsenal forever, won the c
```

In [5]:

```
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 [8]:

```
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:\Users\User\anaconda3\envs\python-dscourse\Lib\site-packages\urllib3\con
nectionpool.py:1056: InsecureRequestWarning: Unverified HTTPS request is b
eing made to host 'www.agc.gov.my'. Adding certificate verification is str
ongly advised. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usag
e.html#ssl-warnings (https://urllib3.readthedocs.io/en/1.26.x/advanced-usa
ge.html#ssl-warnings)
  warnings.warn(
```

Out[8]:

```
1485266
```

In [9]:

```
# 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 [10]:

```
page_content
```

Out[10]:

```
'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 [11]:

```
# %matplotlib inline
```

In [12]:

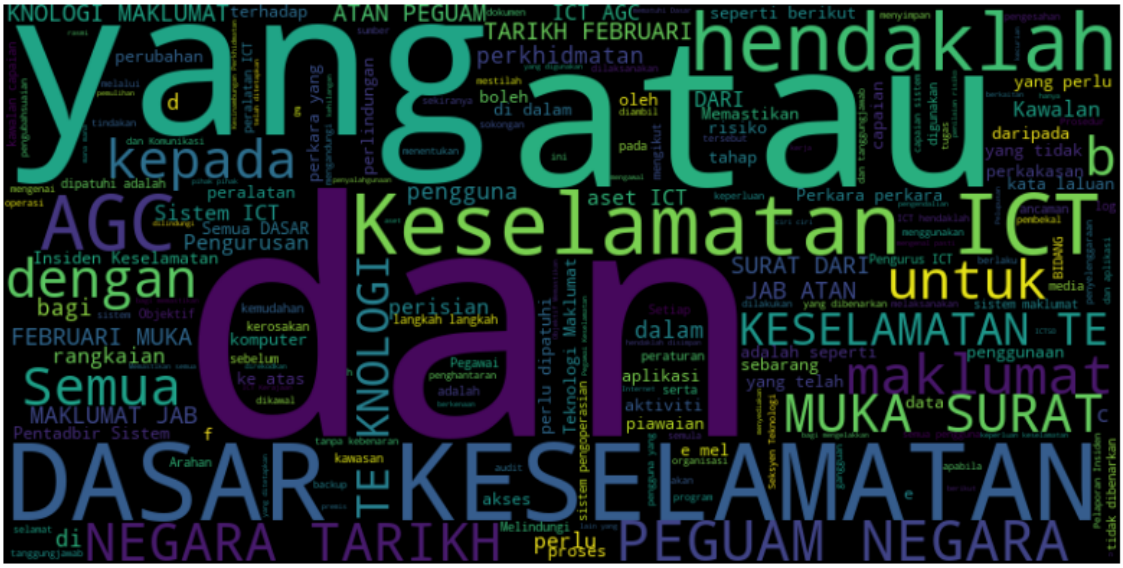
```
from wordcloud import WordCloud, STOPWORDS

# 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 [13]:

```
# 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 [14]:

```
# 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 [15]:

```
# 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[15]:

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

In [16]:

```
# 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[16]:

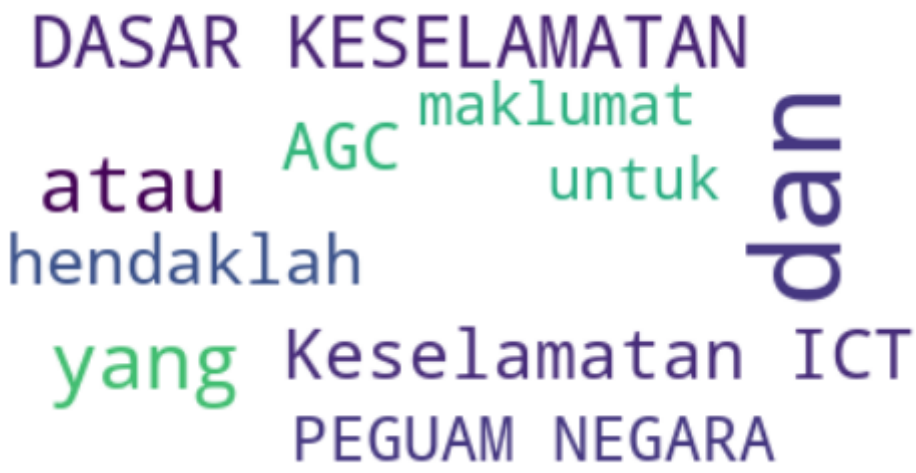
```
<function matplotlib.pyplot.show(close=None, block=None)>
```

In [17]:

```
# 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[17]:

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



In [18]:

```
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[18]:

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

In [19]:

```
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[19]:

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

In [20]:

```
stopwords
```

Out[20]:

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

In [21]:

```
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[21]:

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

Mask

Change the layout

Generate a Numpy grid

In [26]:

```
# 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 [23]:

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

In [49]:

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

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

color= ImageColorGenerator(mask)

wordcloud = WordCloud(width=50,
                      height=50,
                      max_words=10000000,
                      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()
```

In [48]:

```
wordcloud.to_file('arsenal_wordcloud.png')
```

Out[48]:

```
<wordcloud.wordcloud.WordCloud at 0x1caa5598250>
```

Read up

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

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 [50]:

```
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)
```

Download from the source

The source code of word_cloud https://github.com/amueller/word_cloud (https://github.com/amueller/word_cloud)

The Jupyter notebooks https://amueller.github.io/word_cloud/ (https://amueller.github.io/word_cloud/)

Quiz

1. Download pdf from this link: https://huntfish.mdc.mo.gov/sites/default/files/downloads/page/IntroToFishing_2017_v2.pdf (https://huntfish.mdc.mo.gov/sites/default/files/downloads/page/IntroToFishing_2017_v2.pdf)
2. Text Visualization without mask for this text (using WordCloud)(Black and White)
3. Text Visualization with a mask (you can choose your prefered mask)

- Put in the url link of your mask

In [68]:

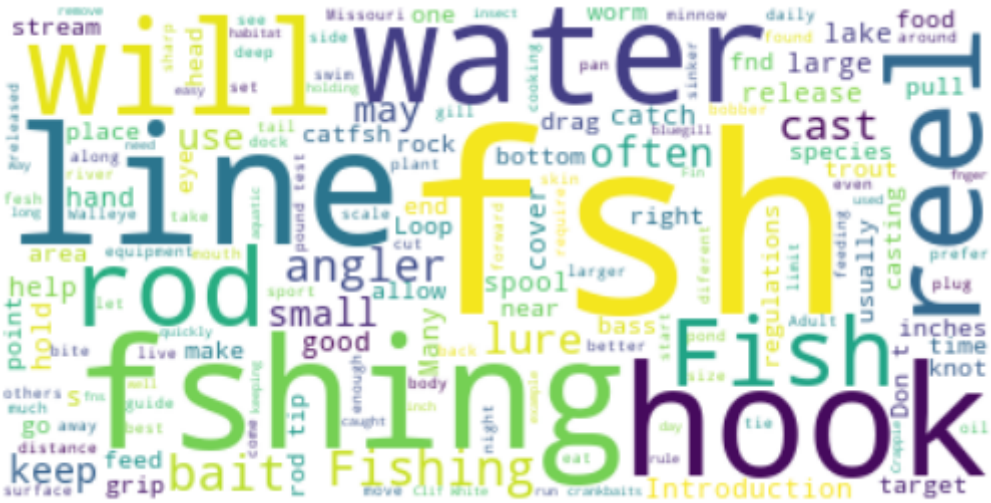
```
from wordcloud import WordCloud, STOPWORDS
from PIL import Image
import urllib
import requests
import numpy as np
import matplotlib.pyplot as plt

text = open('./An Introduction to Fishing.txt', encoding='utf-8').read()
text = text.replace ('  ',' ')
wordcloud = WordCloud().generate(text)
plt.axis("off")
plt.imshow(wordcloud, interpolation='bilinear')
plt.show()
```

In [73]:

```
wordcloud = WordCloud(max_font_size=100, background_color='white').generate(text)
plt.figure()
plt.axis("off")
plt.imshow(wordcloud, interpolation="bilinear")
plt.show
```

Out[73]:

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



In [92]:

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

mask = np.array(Image.open('./p1.png'))

color= ImageColorGenerator(mask)

wordcloud = WordCloud(width=50,
                      height=50,
                      max_words=100000000000,
                      max_font_size=500,
                      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()
```

In [93]:

```
wordcloud.to_file('cramorant_wordcloud.png')
```

Out[93]:

```
<wordcloud.wordcloud.WordCloud at 0x1caa60c8f10>
```

In [ ]:

```

```