Assignment 1

1.1 Every word starts with 'a'



1.2 Every word ends with 'ing' if(re.search(".*ing\$",w)): Code

1.3 Every word contains substring 'pen' or 'men'



1.4 Every word contains substring starts by 'un' or 'pre'



1.5 Every word presents as 2-digit numbers



Assignment 2

```
import nltk
from nltk.tokenize import sent tokenize, word tokenize
from nltk.probability import FreqDist
from nltk.corpus import stopwords
import re
import csv
# Import csv file
data = ""
with open('tweets_all.csv',errors="ignore") as myfile:
    reader = csv.reader(myfile,delimiter = '\n')
    for row in reader:
        data += row[0]
# Find number of sentence
sentence = sent tokenize(data)
print("The number of sentences: ",len(sentence))
# Word tokenize
words = word tokenize(data)
# Create a pattern to find words to search for
pattern = '[a-zA-Z]{3,}'
regex = re.compile(pattern)
search = regex.findall(data)
# Add stop word in english
stop words = set(stopwords.words('english'))
# Filter to remove stop words
wordsFiltered = []
for w in search:
    if w not in stop_words:
       wordsFiltered.append(w)
# Find the most used words
fdist = FreqDist(wordsFiltered)
# Sort the most used words
sortData = sorted(fdist.items(),key=lambda k:k[1] ,reverse=True)
# Show sort of the most used 20 words
for i in range(20):
    print(sortData[i])
```

ผลลัพธ์ที่ได้จากการรัน

```
D:\NLP>c:/python38/python.exe d:/NLP/week2-assignment2.py
The number of sentences: 3362
('world', 331)
('get', 314)
('around', 298)
('enjoy', 297)
('camera', 287)
('bring', 286)
('ready', 285)
('view', 284)
('sunrise', 283)
('sunset', 283)
('Fixed', 283)
('Updated', 283)
('Developed', 283)
('love', 242)
('life', 187)
('Love', 183)
('fan', 177)
('http', 174)
('com', 170)
('God', 124)
```

เพิ่มเติม

เนื่องจาก stop word นั้นไปประกอบไปดังคำต่าง ๆ ตามรูปด้านล่างนั้น ทำให้ผลลัพธ์ที่ได้ไม่ตรงตาม ตัวอย่างผลลัพธ์ที่อย่ assignment

```
stop_words = {'or', 'same', "aren't", 'at', 'over', 'themselves', 'himself', 'having', 'below', 'does', 't', 'it', 'be', 'in', "couldn't", 'wasn', 'm', 'once', 's', 'above', "sh ouldn't", under', 'during', 'up', 'but', 'each', 'weren', 'had', 'our', 'mustn', 'have', 'll', 'nor', 'between', 'why', 'him', "needn't", 've', 'just', 'is', 'what', 'by', 'to', 'd', 'yourselves', "mightn't", 'won', 'again', 'you', 'if', 'out', 'the', 'needn', 'didn't", 'while', 'on', "it's", "you'd', "that'll", "don't", 'not', 'don', 'ma', 'hasn', 'thi s', 'amy', "weren't", "mustn't", 'ours', 'did', 'off', 'your', 'its', 'will', 'yourself', 'my', 'how', 'am', "hadn't", 'such', 'than', 'there', 're', 'as', 'about', 'doing', 'her s', 'before', 'own', 'through', 'doesn', 'after', 'then', 'where', 'yours', 'ourselves', "you'll", 'wasn't", 'should', 'y', 'wouldn't", 'from', 'off, 'she's", 'which', 'l', 'were', 'down', 'now', 'won't", 'ain', 'wouldn', 'theirs', 'shan', 'and', 'all', 'was, 'her', 'he', 'me', 'shouldn', 'isn', 'myself', 'no', 'an', 'few', 'hreself', 'ha ven't", 'very', "doesn't", 'aren', 'until', 'who', 'o', 'haven', "isn't", 'for', 'these', 'we', 'she', 'do', "should'we", 'some', 'been', 'they', 'both', 'has', 'further', 'his', 'when', 'whom', 'here', 'so', 'that', 'couldn', 'more', 'didn'}, 'when', 'whom', 'here', 'so', 'that', 'couldn', 'more', 'didn'}, 'when', 'whom', 'here', 'so', 'that', 'couldn', 'more', 'didn'}
```