CS practice project report template

Project Name: The analysis based on web Spyder and word statistic for lyric preferences of some Japanese musicians

Your Name: Daniel Chu

# Problem definition

## What is the project about? (What is it? Why you want to do it? Possible applications)

Write python program to lyric data from website and do the statistics.

To find the frequency of the usage of particular words

In order to find the preferable theme and preferable word usage of a particular musician

In order to have better understanding of a musician’s artwork, give reference and concept to the study of lyric writing

## 1.2 What are the success criteria? (What are your goals that will guide you through design and implementation and can be evaluated)

1. lyrics can be successively abstracted from websites

2. lyrics can be stored as TXT files, the file name of each TXT file should be the same as the music name it stored

3. read each TXT file and statistic the number of each word present

4. data visualization

5. analysis & conclution

# Design overview

How to solve the problem in part 1? You need to design your software system, design about the input, process, and output parts.

import urllib

from bs4 import BeautifulSoup as bs

#input

resp = urllib.request.urlopen('http://blog.sina.com.cn/cqryg')

html\_data = resp.read().decode('utf-8')

soup = bs(html\_data, 'html.parser')

#/////////////////////////

#process

music\_info = soup.find\_all('div', class\_ = 'blog\_title')

counter = 0

list\_info = []

for i in music\_info:

music\_info\_1 = music\_info[counter].find\_all('a', target = '\_blank')

counter = counter + 1

for item in music\_info\_1:

dic\_info = {}

operat = str(item).split()

ele = dic\_info['Title: '] = operat[1][6:-1]

list\_info.append(dic\_info)

#/////////////////////////

def Initialize\_Soup(Link):

Resp = urllib.request.urlopen(link)

Html\_Data = Resp.read().decode('utf-8')

Soup = bs(Html\_Data, 'html.parser')

return Soup

def Find\_Style(Soup): #Soup: list

for i in Soup:

Counter = 0

Line = str(i) if i != None else ''

if Line != '':

for c in Line:

if Line[Counter: Counter + 9] == 'font-size':

list1.append(i)

Counter = Counter + 1

if Counter == len(Line) - 9:

break

def Abstract\_Text(Soup\_Music\_Script\_List):

for i in Soup\_Music\_Script\_List:

Text\_Sum = i.text

return Text\_Sum

def Abstract\_Title(Soup):

Title = Soup.find('h2').text if Soup.find('h2') != None else 'unknown'

return Title

def Write\_TextFile(Text, FileName):

Text = str(Text)

File\_Holder = open(str(FileName) + '.txt', 'w', errors='ignore')

File\_Holder.write(Text)

File\_Holder.close()

for item in list\_info:

#process

list1 = []

link = item['Title: ']

soup = Initialize\_Soup(link)

music\_script = soup.find\_all('div')

Find\_Style(music\_script)

#output

Write\_TextFile(Abstract\_Text(list1), Abstract\_Title(soup))

# Implementation

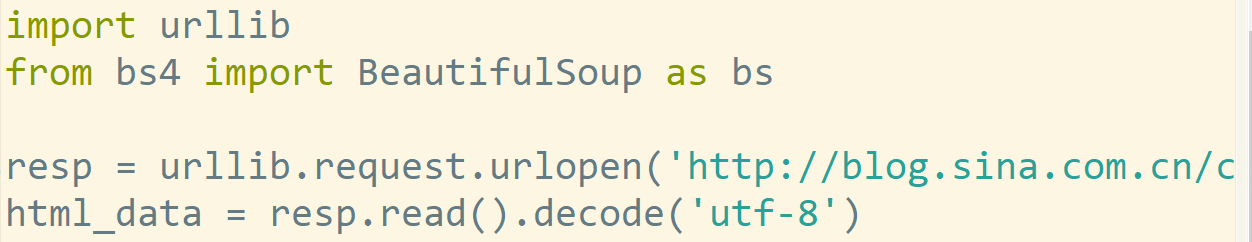
List the key techniques used in your software, like arrays, searching, sorting, dictionaries, python modules (third-party, like matplotlib, numpy, scipy, sklearn etc.)

Tech #1: urllib

Why you use tech 1 in your project?

To decode and encode the data

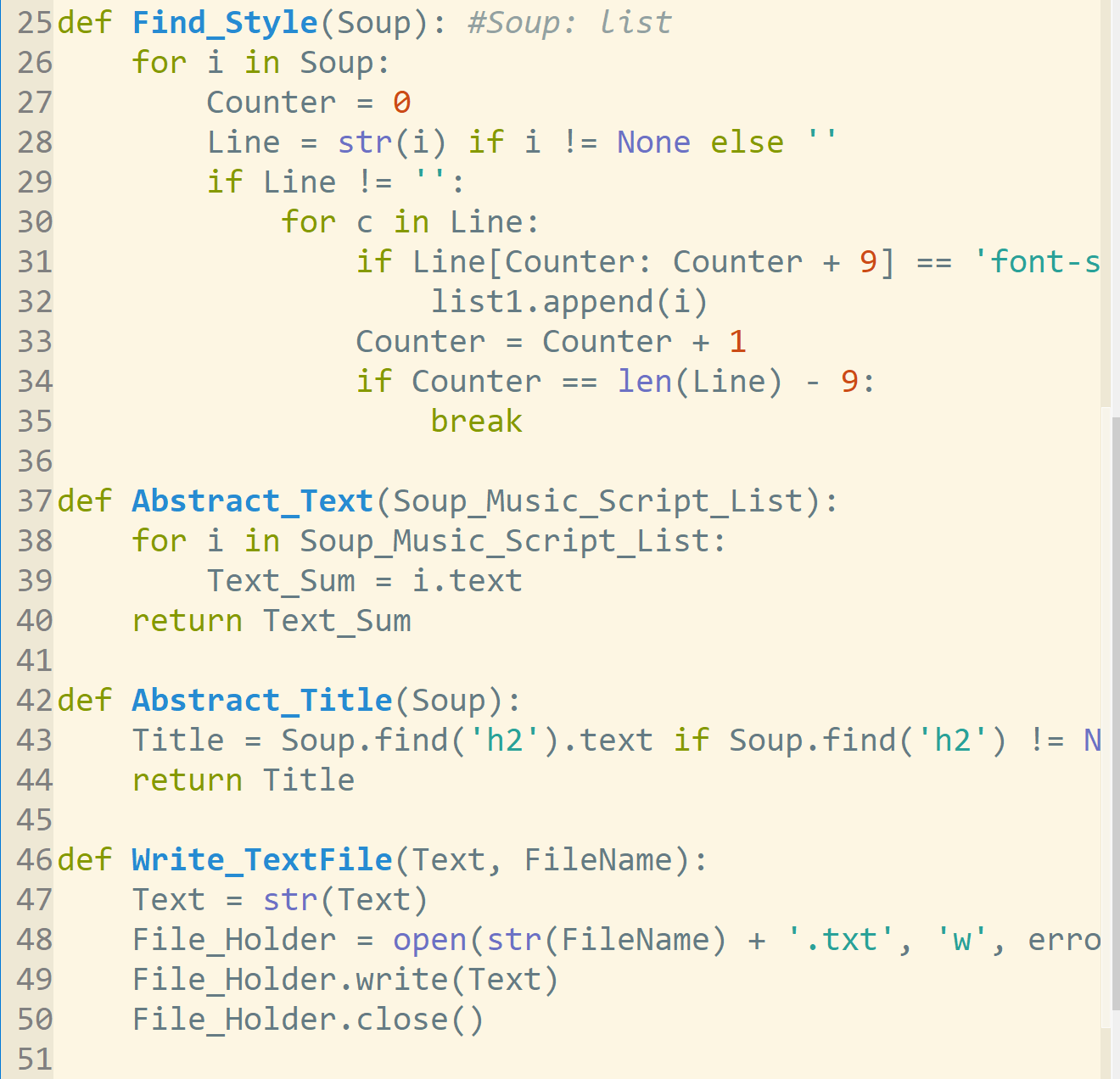
How do you use it? Put your code screenshots here.

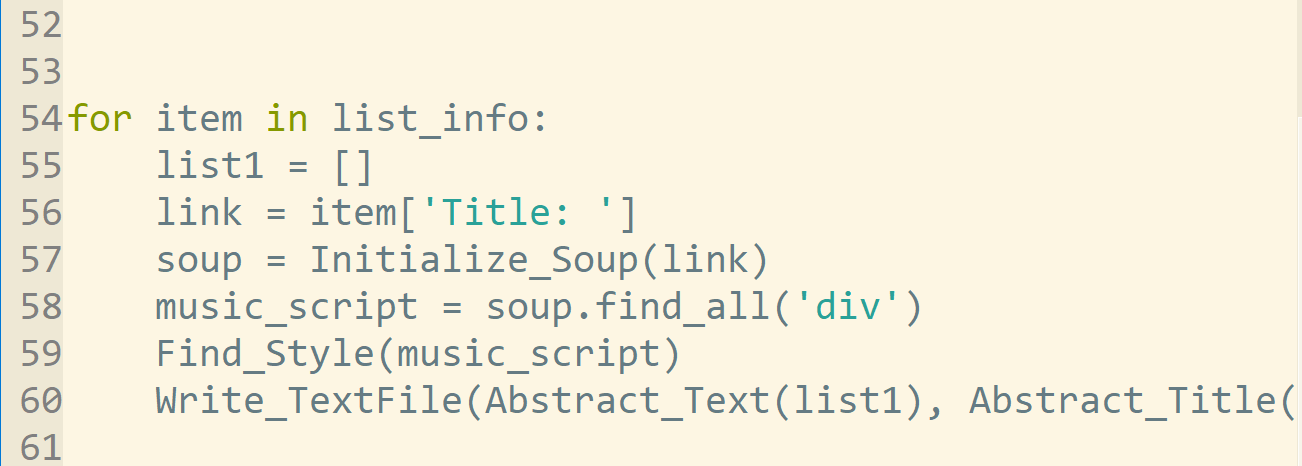


Tech2: BeautifulSoup4

To abstract lyric from web site.







# Results

Show the outcomes/results of your projects



# Evaluation and conclusion

Abstraction success

# References

List all the materials you referred to.

<https://segmentfault.com/a/1190000010473819>

<https://ask.hellobi.com/blog/yuguiyang1990/9445>

<https://github.com/wudithu08/icc-AL-CS-2021/blob/master/Assessments/files/resources-textmining.zip>

<https://github.com/wudithu08/icc-AL-CS-2021/blob/master/Assessments/files/douban-movie.zip>

<https://www.cnblogs.com/delav/p/7845539.html>

<https://www.nippon.com/en/behind/l10797/japan-may-defer-announcement-of-new-era-name.html>