CSGE601020
Dasar-Dasar Pemrograman 1
(Foundations of Programming 1)

Programming Assignment 2: Word Clouds

LAST DAY for uploading the result of your work to SCeLE: Fri 8 Oct 2021 (11:55 PM SCeLE time).

Don’t forget to write enough comments in your source code.

Please contact your TA (teaching assistant) for giving a demo of your work as soon as possible. The TA will give you a mark after the demo.

Please start working on this assignment immediately. If you have any questions, please ask the TA or the teacher.

Marking scheme:

60 % correctness
30 % explanation in demo session
10 % program documentation (comments, neatness)

Task Description

Purpose

This assignment will give you useful experience with functions, strings, lists, text processing, as well as file I/O.

Problem

A common element seen on web pages these days are word clouds or tag clouds of a document. A word cloud is a visual representation of frequency of words, where more frequent words are represented in larger font.

You are going to analyze a text document given by the user and create a word cloud for it, where the size of the font in the cloud indicates the frequency of the words.

Background

You are provided with a number of elements to help with this assignment:

- text documents for processing: CommencementSpeechByGates2014.txt, JokoWidodoSpeechAPEC2014.txt

- a file containing stop words.
- some functions for creating an HTML (hypertext markup language) file which can be displayed by a standard browser.

Stop Words

Not all words are worth counting. 'a', 'the', 'was', 'and', etc. are just junk. A list of such words is provided in the file **'stopWords.txt'.** Each line has a single word. No word in the stop word list should be counted in the word cloud.

Functions

Three functions and an example are provided in **htmlFunctions.py**. Use the functions in your program. That file contains:

- `make_HTML_word(word, count, high, low)`: This function takes a `word` and wraps it in a font tag with a specific size and a random color — returning that string whose `fontsize` is between `htmlBig` and `htmlLittle` (two local variables in the function that you can change to be whatever you like). The parameters are:
    - `word` (string), the word to be wrapped,
    - `count`, how many times it occurred in the input document,
    - `high`, the highest word count in the input document,
    - `low`, the lowest word count in the input document.

  The function returns the word as a HTML-formatted string.

- `make_HTML_box(body)`. This function takes a single string named `body` that contains all the font-wrapped words from `make_HTML_word` and places them in an HTML box to be displayed. It returns a string which is the HTML code for the box.
- `print_HTML_file(body,title)`. Takes the `body` (string) returned from `make_HTML_box`, wraps a standard HTML web page tags around it, and creates a file. The string `title` is used in the HTML. The title is also used as the file name with an '.html' suffix. The file is created and written to secondary storage.

Program Specification

- Prompt the user for the text file to be processed.
- Print the top 56 counts (word and count pairs), in the format of 14 rows by 4 columns. Sort the words by count (frequency) for printing. (For the next step you will sort the words alphabetically for the HTML file.)
- Generate an HTML file (with suffix .html) using the provided functions to generate a word-cloud file with the top 56 words. For the word cloud sort the top 56 words alphabetically. The word cloud looks more interesting that way. You can view the files in your favorite browser.

Hints:

There are a couple of problems here. Think about each one before you start to program.

1. You have to read in the file and separate the contents into words.

2. Once you have the words separated, you must remove all stop words (using the provided file “stopWords.txt”). Also, remember to remove punctuation from words, just because a word comes at the end of a sentence and has a period at the end of it doesn’t make it a different word. (Importing `string` and using `string.punctuation` is a useful way to specify punctuation.)

3. You must then count the frequency of each word collected. Use the list data structure or the dictionary data structure.

4. Once you have a complete list, sort the list. Put count first in the tuple because sorting (using either `sort` or `sorted`) will sort on the first item.

5. Use the provided functions to turn the words and counts into an HTML page (see the example in the file **htmlFunctions.py**).

Happy Programming! ’Met ngoding!

L. Y. Stefanus

Sample Ouput:

1. First is what is printed to the IDLE shell (note: sorted by count).

```
Program to create word cloud from a text file
---------------------------------------------
The result is stored as an HTML file,
which can be displayed in a web browser.

Please enter the file name: CommencementSpeechByGates2014.txt

CommencementSpeechByGates2014.txt :
56 words in frequency order as (count:word) pairs

 26:people       17:optimism       12:world          12:melinda
 11:stanford     11:make            9:bill            8:poor
  8:innovation    8:computers       7:leave           7:knew
  7:kids          6:years           6:workers         6:time
  6:talk          6:sex             6:problems        6:empathy
  6:cure          6:change          6:center          6:aids
  5:women         5:wanted          5:trip            5:technology
  5:soweto        5:south           5:poverty         5:patients
  5:made          5:lives           5:hospital        5:future
  5:foundation    5:day             5:children        4:united
  4:things        4:tb              4:suffering       4:states
  4:software      4:power           4:pessimists      4:microsoft
  4:luck          4:home            4:helped          4:heart
  4:gates         4:empower         4:community       4:africa

Please type Enter to exit ...
```

2. Next is the generated HTML file: CommencementSpeechByGates2014.txt.html.

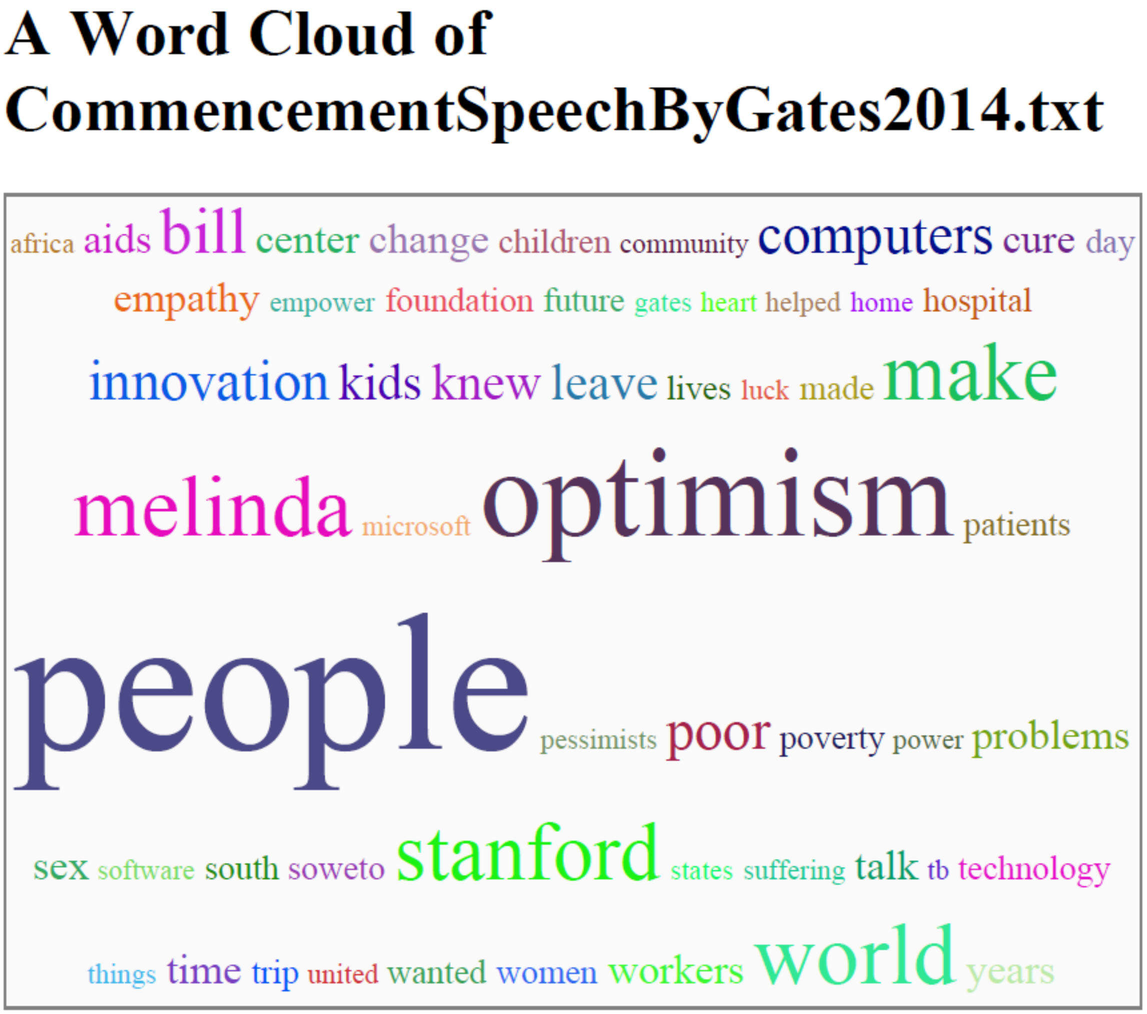
```
    <html> <head>
    <title>A Word Cloud of CommencementSpeechByGates2014.txt</title>
    </head>

    <body>
    <h1>A Word Cloud of CommencementSpeechByGates2014.txt</h1>
<div style="
    width: 560px;
    background-color: rgb(250,250,250);
    border: 1px grey solid;
    text-align: center" ><span style="color: rgb(182, 125, 40); font-
size:14px;">africa</span> <span style="color: rgb(201, 22, 200); font-
size:21px;">aids</span> <span style="color: rgb(200, 35, 215); font-
size:32px;">bill</span> <span style="color: rgb(37, 177, 97); font-
size:21px;">center</span> <span style="color: rgb(153, 117, 175); font-
size:21px;">change</span> <span style="color: rgb(170, 85, 107); font-
size:17px;">children</span> <span style="color: rgb(93, 26, 68); font-
size:14px;">community</span> <span style="color: rgb(3, 15, 138); font-
size:28px;">computers</span> <span style="color: rgb(111, 32, 140); font-
size:21px;">cure</span> <span style="color: rgb(131, 110, 173); font-
size:17px;">day</span> <span style="color: rgb(234, 102, 19); font-
size:21px;">empathy</span> <span style="color: rgb(57, 177, 156); font-
size:14px;">empower</span> <span style="color: rgb(233, 74, 91); font-
size:17px;">foundation</span> <span style="color: rgb(59, 172, 102); font-
size:17px;">future</span> <span style="color: rgb(34, 232, 145); font-
size:14px;">gates</span> <span style="color: rgb(79, 255, 8); font-
size:14px;">heart</span> <span style="color: rgb(166, 125, 88); font-
size:14px;">helped</span> <span style="color: rgb(165, 7, 249); font-
size:14px;">home</span> <span style="color: rgb(194, 84, 0); font-
size:17px;">hospital</span> <span style="color: rgb(9, 91, 228); font-
size:28px;">innovation</span> <span style="color: rgb(76, 0, 178); font-
size:25px;">kids</span> <span style="color: rgb(162, 27, 197); font-
size:25px;">knew</span> <span style="color: rgb(39, 122, 167); font-
size:25px;">leave</span> <span style="color: rgb(35, 102, 3); font-
size:17px;">lives</span> <span style="color: rgb(226, 91, 65); font-
size:14px;">luck</span> <span style="color: rgb(169, 155, 7); font-
size:17px;">made</span> <span style="color: rgb(26, 192, 91); font-
size:40px;">make</span> <span style="color: rgb(231, 11, 190); font-
size:43px;">melinda</span> <span style="color: rgb(243, 167, 105); font-
size:14px;">microsoft</span> <span style="color: rgb(85, 50, 90); font-
size:62px;">optimism</span> <span style="color: rgb(131, 109, 33); font-
size:17px;">patients</span> <span style="color: rgb(76, 73, 136); font-
size:96px;">people</span> <span style="color: rgb(114, 147, 98); font-
size:14px;">pessimists</span> <span style="color: rgb(166, 33, 73); font-
size:28px;">poor</span> <span style="color: rgb(33, 29, 89); font-
size:17px;">poverty</span> <span style="color: rgb(80, 103, 61); font-
size:14px;">power</span> <span style="color: rgb(113, 164, 17); font-
size:21px;">problems</span> <span style="color: rgb(51, 170, 96); font-
size:21px;">sex</span> <span style="color: rgb(119, 220, 96); font-
size:14px;">software</span> <span style="color: rgb(31, 140, 9); font-
size:17px;">south</span> <span style="color: rgb(153, 61, 179); font-
size:17px;">soweto</span> <span style="color: rgb(28, 241, 20); font-
size:40px;">stanford</span> <span style="color: rgb(13, 249, 105); font-
size:14px;">states</span> <span style="color: rgb(25, 200, 145); font-
size:14px;">suffering</span> <span style="color: rgb(3, 157, 102); font-
size:21px;">talk</span> <span style="color: rgb(90, 37, 203); font-
size:14px;">tb</span> <span style="color: rgb(230, 17, 200); font-
size:17px;">technology</span> <span style="color: rgb(62, 183, 235); font-
size:14px;">things</span> <span style="color: rgb(120, 63, 192); font-
size:21px;">time</span> <span style="color: rgb(3, 83, 249); font-
size:17px;">trip</span> <span style="color: rgb(208, 51, 54); font-
size:14px;">united</span> <span style="color: rgb(53, 159, 84); font-
size:17px;">wanted</span> <span style="color: rgb(44, 103, 241); font-
```

```
size:17px;">women</span> <span style="color: rgb(57, 247, 49); font-
size:21px;">workers</span> <span style="color: rgb(48, 231, 147); font-
size:43px;">world</span> <span style="color: rgb(188, 234, 169); font-
size:21px;">years</span></div>


    </body> </html>
```

Here is what the CommencementSpeechByGates2014.txt.html file looks like in a browser (note: the words are sorted alphabetically).



A Word Cloud of CommencementSpeechByGates2014.txt

africa aids bill center change children community computers cure day empathy empower foundation future gates heart helped home hospital innovation kids knew leave lives luck made make melinda microsoft optimism patients people pessimists poor poverty power problems sex software south soweto stanford states suffering talk tb technology things time trip united wanted women workers world years