Data Structure

Fall 2020, programming Assignment #1 Note

Due Date: 23:55 on Monday, December 7th, 2020

Submission: Place all files in a folder and zip it. Both the zip file and the folder inside should be named with your student ID (e.g., B123456789.zip with a folder named B123456789 in it). Upload the zip file via CEIBA. The folder should contain the following:

* You will need to submit 26 files plus files for your source code. So please pay close attention to the instructions below. The names of these 26 files are already defined. Any files with unrecognizable names will be discarded (and therefore might have adverse effects on your grades).

i) **Source code files**: You can use the programming language of your choice. Please include all the source code needed to run / compile your program.

ii) **Readme.pdf**: The file named "Readme.pdf" is a simple report that describes how your solution is implemented (e.g., data structure, algorithm, ...). Please also include a simple performance analysis on time / space complexity. Optionally, you can include simple explanations on how your code is structured. This file should not exceed two pages.

iii) **Page Rank files**: You need to run your program with the following combinations of arguments: d = 0.25, 0.45, 0.65, 0.85 and DIFF = 0.100, 0.010, 0.001. For each of the 12 combinations, create a text file with a list of the pages in page rank order (highest to lowest), for each page listing its outbranching (how many pages it connects to) and the page rank for that page (to 8 significant digits). Please see page 3 of prog1.pdf (under the section "Page Rank list") for the detailed format and example. The name of the page rank file is defined as **pr_xx_yyy.txt**, where xx is $d \times 100$ and yyy is DIFF $\times$ 1000. For example, for d=0.25 and DIFF=0.010, the file name should be pr_25_010.txt.

iv) **Reverse index file**: Output the reverse index into a text file named "**reverseindex.txt**". Please see page 3 of prog1.pdf (under the section "Reverse index") for details on file format and an example. Note: please use ASCII order, i.e., the word "NASA" should be listed before the word "dog".

v) **Search Engine**: Each line of the file list.txt contains a list of words. With respect to each of the above combinations of d (say, 0.xx) and DIFF (say, 0.yyy), perform search on the lists of words line by line and output the top ten pages (sorted by page rank) that contain them (if less than 10,

then list all of them) into file "result_xx_yyy.txt". For instance, for d=0.25, DIFF=0.010, output the file "result_25_010.txt". When the line read in contains a single word, just output the pages for that word. When the line read in contains multiple words, output two lists - the top ten pages with all the words (AND semantics) and the top ten pages with any of the words (OR semantics). If none of the pages contain the read-in words, output the word "none" into the file. This is essentially the output from your search engine as specified in page 3 of prog1.pdf (under the section "Search engine"), with the addition of the "none" for when no pages were found. Following is an example on how the input (list.txt) and your output files (result_xx_yyy.txt) look like. **_Note_**: *Uppercase letters and lowercase letters should be treated differently. Space character and newline character are different from each other.*

Example:

Content of file "list.txt":

```
Dog
cat walk
NASA
men's book three ...
```

Content of file "result_xx_yyy.txt" (your output file):

```
page1000 page200 page2 page4 page5 page59 page3 page11 page10 page13
AND none
OR page15 page200 page7 page4 page43 page9 page3 page11 page77 page13
none
AND page200 page7 page9
OR page200 page23 page7 page9 page77 page13
...
```

If you have any questions regarding the requirement specified in this file or any of the materials related to this assignment listed at the course website, pleases do not hesitate to contact any one of the three TA's via emails:

王璈	d07921012@ntu.edu.tw

王思涵	stellawang18@gmail.com

余若萱	q8611027@gmail.com