Bid Data Mining Homework 5

Team Member:

SID: 109598033 SID: 109598001

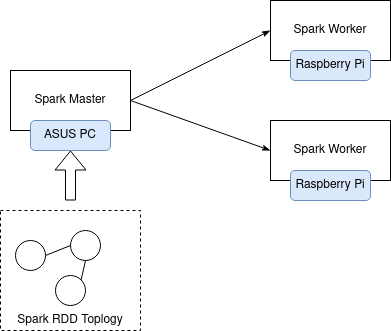
**Spark Platform:**

The platform consists of:  
1. Raspberry Pi 4 Model B x2

* OS: Linux Ubuntu 20.04 Server
* CPU architecture: aarch64
* RAM: 8GB
* CPU: Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz
* Number of CPU: 4C (CPU) 1T (Thread Per CPU)

2. Asus-vivobook notebook

* OS: Linux Ubuntu 20.04 LTS
* CPU architecture: x86\_64
* RAM: 8GB
* CPU: Intel(R) Core(TM) i3-8130U CPU @ 2.20GHz
* Number of CPU: 4C (CPU) 2T (Thread Per CPU)

The simple architecture of spark cluster:

**Task arrangement for Team:   
SID: 109598033  
1. Q1  
2. Q2  
SID: 109598001  
1. Q3  
2. Document**

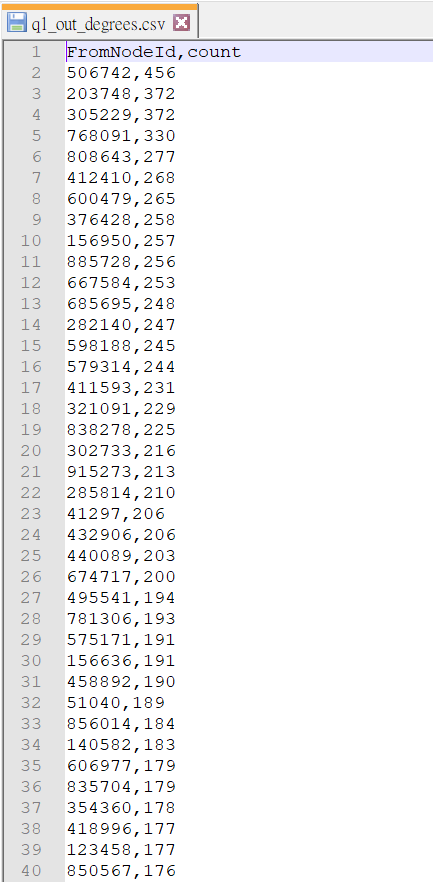
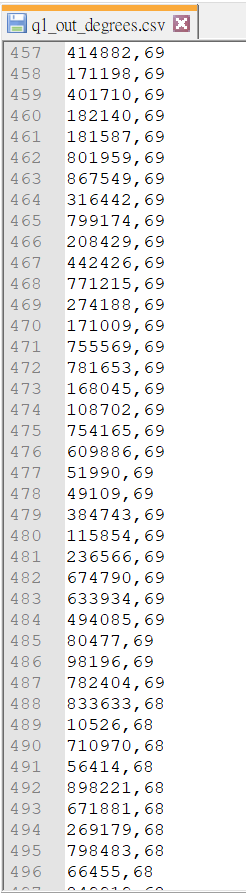
**The description of tree directory:**

Explain where the files put into and how it does.

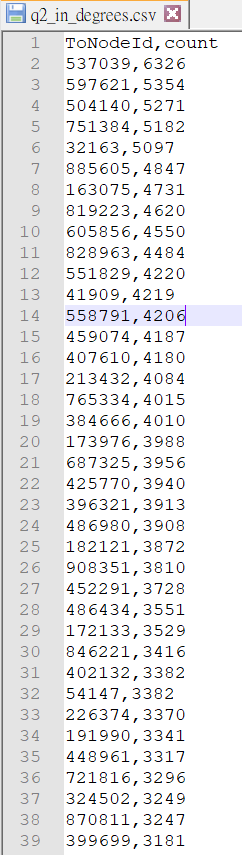
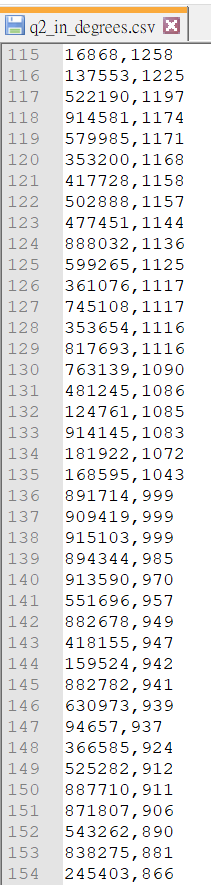
* **README file explains in detail for what the steps of spark implementation in this homework.**
* **‘outputs’ directory shows the result of the homework.**
* **park project is ‘109598033\_109598001\_hw5.ipynb’ which codes by Python.**

**The generated output:**

**Q1—web pages with the number of outlinks, sorted in descending order**

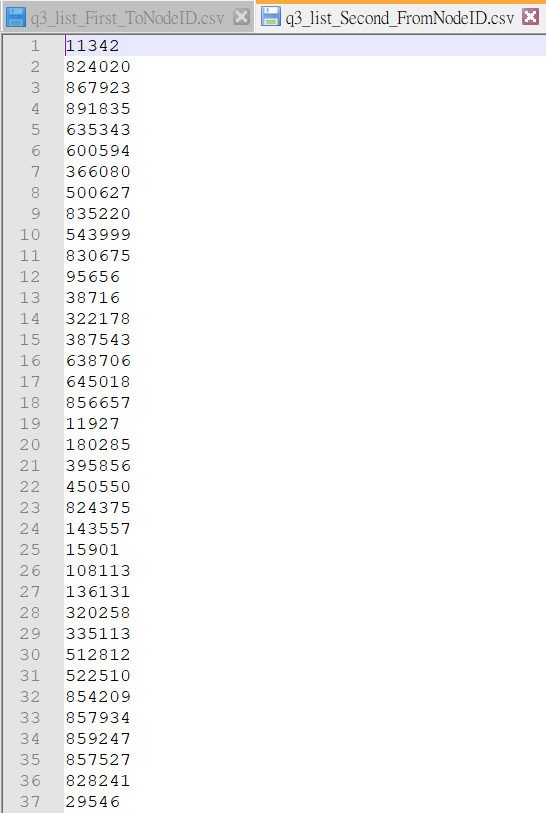
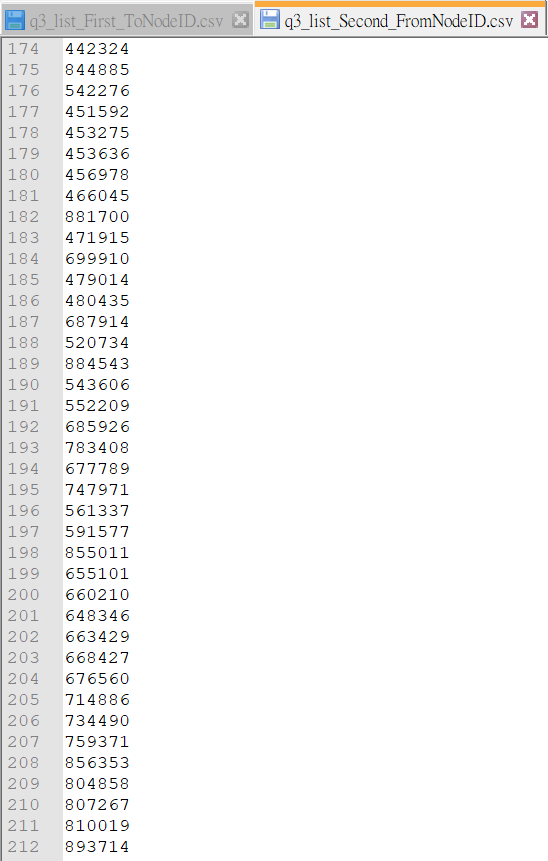
 

**Q2—inlink distribution of the top linked web pages, sorted in descending order**

**Q3—output the list of nodes that v points to, and the list of nodes that points to v**

FromNodes:

ToNodes:

