Hadoop Assignment - 3

You are a system engineer planning to deploy a page rank algorithm. You crawled the web and collected the following data:

Source	***Linked To***
WS1	WS3,WS5
WS2	WS1
WS3	WS4

The term “WS” stands for WebSite. So from the table, one can infer the following:

- **WS2 is 1 Click away from WS1**
    - Please note that, from WS1 you cannot visit WS2 so the opposite is not always true
- **WS1 is 2 Clicks away from WS4**
    - WS1 – WS3 – WS4

Click Distance: The number of clicks required to visit a website from the source website.

Task:
- For a given dataset and X, the code should be able to output all the records that are having the Click Distance as X.

- Eg:hadoop jar YOURCODE.JAR YOURMAIN **X** /INPUT/PATH /OUTPUT/PATH

- If X is set to 2, for the above table, the output should be as follows:

SOURCE – DESTINATION(LIST OF CONNECTING NODES)

For Eg: WS1-WS4,(WS3)

- If for a given value of X if there are no values present the output can be empty
- Assume that each record has only one entry

For Eg: First record is stored as:

- WS1,WS3
- WS1,WS5

Your code will be executed on the test data