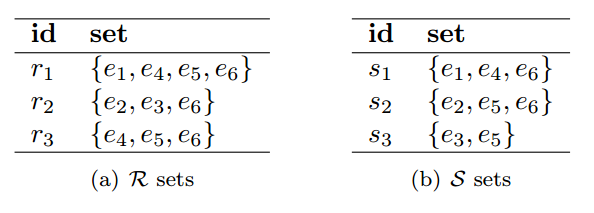
## COMP9313 2022T3 Project 3 (22 marks)

## Similarity Join Using Pyspark

### Problem Definition:

Given two collections of records R and S, a similarity function **sim**(., .), and a threshold τ, the set similarity join between R and S, is to find all record pairs r (from R) and s (from S), such that **sim**(r, s) >= τ. We compute **sim**(., .) using the Jaccard similarity in this project.



Given the above example, and set τ=0.5, the results are (r1, s1) (similarity 0.75), (r2, s2) (similarity 0.5), (r3, s1) (similarity 0.5), (r3, s2) (similarity 0.5).

### Input files:

Each set is stored in one text file, and each line is in format of: “RecordId list<ElementId>”. Two example input files are as below (elements are separated by space):

|  |  |
| --- | --- |
| File1 | File2 |
| 0 e1 e4 e5 e6  1 e2 e3 e6  2 e4 e5 e6 | 0 e1 e4 e6  1 e2 e5 e6  2 e3 e5 |

Another small test data set can be downloaded at (τ=0.1): <https://webcms3.cse.unsw.edu.au/COMP9313/22T3/resources/82369>

**Note that it is possible that one element appears multiple times in a record. You need to convert the record to a set first to compute the Jaccard similarity.**

### Output:

The output file contains the similar pairs together with their similarity scores. Each line is in format of “(RecordId1,RecordId2)**\t**Similarity” (RecordId1 is from the first file and RecordId2 is from the second file). **Round the similarities to six decimal places.** The pairs are sorted in ascending order by the first record and then the second. Given the example input data, the output file is like:

|  |
| --- |
| (0,0)**\t**0.75  (1,1)**\t**0.5  (2,0)**\t**0.5  (2,1)**\t**0.5 |

### Code format:

Please name your python file as “project3.py” and compress it in a package named “zID\_proj3.zip” (e.g. z5123456\_proj3.zip).

### Command of running your code:

Your program should take four parameters: the two input files, the output folder, and the similarity threshold τ. We will use the following command to run your code:

**$ spark-submit project3.py file1 file2 tau output**

Please ensure that the code you submit can be compiled. Any solution that has compilation errors will receive no more than 6 marks.

## Documentation and code readability

Your source code will be inspected and marked based on readability and ease of understanding. The efficiency and scalability of this project is very important and will be evaluated as well. Below is an indicative marking scheme:

|  |
| --- |
| Result correctness: 12 |
| Efficiency and Scalability: 8 |
| Code structure, Readability, and Documentation: 2 |

## Submission:

Deadline: Mon 21st November 11:59:59 PM

You can submit through Moodle. If you submit your assignment more than once, the last submission will replace the previous one. To prove successful submission, please take a screenshot as assignment submission instructions show and keep it by yourself. If you have any problems in submissions, please email to [siqing.li@unsw.edu.au](mailto:siqing.li@unsw.edu.au).

## Late submission penalty

5% reduction of your marks for up to 5 days

## Plagiarism:

The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such an offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be examined manually.   
  
Relevant scholarship authorities will be informed if students holding scholarships are involved in an incident of plagiarism or other misconduct.   
  
Do not provide or show your assignment work to any other person - apart from the teaching staff of this subject. If you knowingly provide or show your assignment work to another person for any reason, and work derived from it is submitted you may be penalized, even if the work was submitted without your knowledge or consent.