Scalable and Distributed Computing
HCMC International University
Assoc. Prof. Tran Manh Ha

Course: IT139IU
Start Date: 2021-11-20
End Date: 2021-12-25

Mini-Project

(*Group submission*)

Specification

This mini-project aims to provide solutions for large-scale problems. Through doing the mini-project, students learn textual data processing techniques that can deal with a large data corpus. Students form groups of 3 students; each group works with an assigned dataset, then submits reports with deliverables by the deadline. Groups confirm members and problems by 20/11/2021.

Problem 1.1: *Latent Semantic Analysis* (100 points)

Each group uses the Latent Semantic Analysis method and the learned lessons to seek similar documents from the textual dataset to the provided queries. The field-specific dataset is provided by the Instructor. The solution must include text data pre-processing, term-document matrix generation, and singular value decomposition.

The assigned dataset is large, then generated matrices are also large, computation cost is therefore high. The larger dataset a group successfully solves, the higher score the group achieves. Groups must select the different citation datasets in CSV/XML format (starting with a file and then extending to other files): [Download]. Table 1 assigns datasets to groups.

Table 1: Dataset assignment

#	Group ID	Dataset ID	Note
1	Group01	Dataset.part01, part12, part23, ...	
2	Group02	Dataset.part02, part13, part24, ...	
3	Group03	Dataset.part03, part14, part25, ...	
4	Group04	Dataset.part04, part15, part26, ...	
5	Group05	Dataset.part05, part16, part27, ...	
6	Group06	Dataset.part06, part17, part28, ...	
7	Group07	Dataset.part07, part18, part29, ...	
8	Group08	Dataset.part08, part19, part30, ...	
9	Group09	Dataset.part09, part20, part31, ...	
10	Group10	Dataset.part10, part21, part32, ...	
11	Group11	Dataset.part11, part22, part33, ...	

Deliverables Requirements

- D1: A report describing data pre-processing, data processing with algorithm and implementation, data post-processing. Note that the solutions and results including design, algorithm, implementation, ... must be presented in diagrams, flowcharts, pseudo-code, ... while code snippets should be included in the appendices. (70%)
- D2: Each group presents the solutions and results, and specifies members' contribution (30%)