Mansoura University
Faculty of Computers and Information
Department of Computer Science

B.SC. PROJECT PROPOSAL (2018/2019)

Arabic Title

English Title

Deduplication

Submitted by:

Student Name	Student Email
Maha Abdulkader Elmenshawy	Maha.elmenshawy97@gmail.com

PROJECT ABSTRACT:

Data Deduplication is a capacity optimization technology that is being used to dramatically improve storage efficiency.

We define deduplication to mean an operation that matches rows and removes duplicates within one dataset and record linkage as an operation that combines two datasets.

PROJECT OBJECTIVES:

Problem Statement….

Identifying for any dataset the set of records which refer to unique entities.

This problem is known by the varied names of deduplication, identity uncertainty and record linkage.

Project idea …

One common approach is to cast the deduplication problem as a classification problem. Consider the set of record-pairs, and classify them as either "unique" or "not-unique".

*Related work….

Data deduplication technology is being widely used by organizations and companies in order to reduce storage needs and costs, as well as the amount of energy and processing power. Deduplication is usually implemented as part of a storage or backup system.

Example…

*deduplication on Encrypted Big Data in cloud

* Email deduplication

Motivations….

* Data deduplication only store the unique instance of data.

*The redundant data get eliminated and replaced with a pointer to the unique data copy.

*The benefit of data deduplication is obvious.

*Eliminating redundant data can significantly shrink storage requirements and lower storage costs.

*Deduplication also improve the network bandwidth efficiency and save the processing power.

METHODOLOGY:

Datasets…

***Cora.**

***Second String.**

***steam video games.**

***20 years of games.**

Algorithms…

***SVM (Support Vector Machine).**

***Naive Bayes.**

***Deduplication Algorithm.**

SCHEDULING PHASES:

From	To	Activity
7 Nov 2018	14 Nov 2018	Proposal
14 Nov 2018	21 Nov 2018	Dataset & Method
21 Nov 2018	28 Nov 2018	Training & Testing
28 Nov 2018	5 Des 2018	Accuracy

REFERENCES:

https://pdfs.semanticscholar.org/2b14/38962a1a25bae43768ab74e63dfccd9859af.pdf

http://www.dynamicsolutions.com/assets/pdfs/How_DeDupe_Works.pdf

https://www.kaggle.com/rtatman/datacleaning-challenge-deduplication/notebook

https://link.springer.com/chapter/10.1007/978-3-642-20630-6_3

https://recordlinkage.readthedocs.io/en/latest/notebooks/data_deduplication.html

http://www.cs.utexas.edu/users/ml/riddle/data.html

https://d-nb.info/1037457633/34