**LAB 4 (Group 1)**

**Bikash Giri (101575097)**

**Chia-Wei Chang (101570243)**

**Diparshan Bhattarai (101494737)**

**Hsi-Teng Liu (101576074)**

**Gavriel Kirichenko (101119609)**

**Abdul-Rasaq Omisesan (101571156)**

**Callum Arul (101585383)**

**Friba Hussainyar (101591222)**

**Q1: Calculate the bag of words (without TF-IDF, just word counts)  
matrix for the following 5 phrases.**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Phrases/Document** | college | student | airport | taxi | driver | toronto | Art | museum | life | machine | learning |
| college student | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| airport taxi driver | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Toronto art museum | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| Life art | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| Machine learning student | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

* We have 5 rows and 11 columns

**Q2: Calculate the IDF for the words “college” and “art”.**

IDF for college: log(5/1) = 0.689

IDF for art: log(5/2)=0.397``