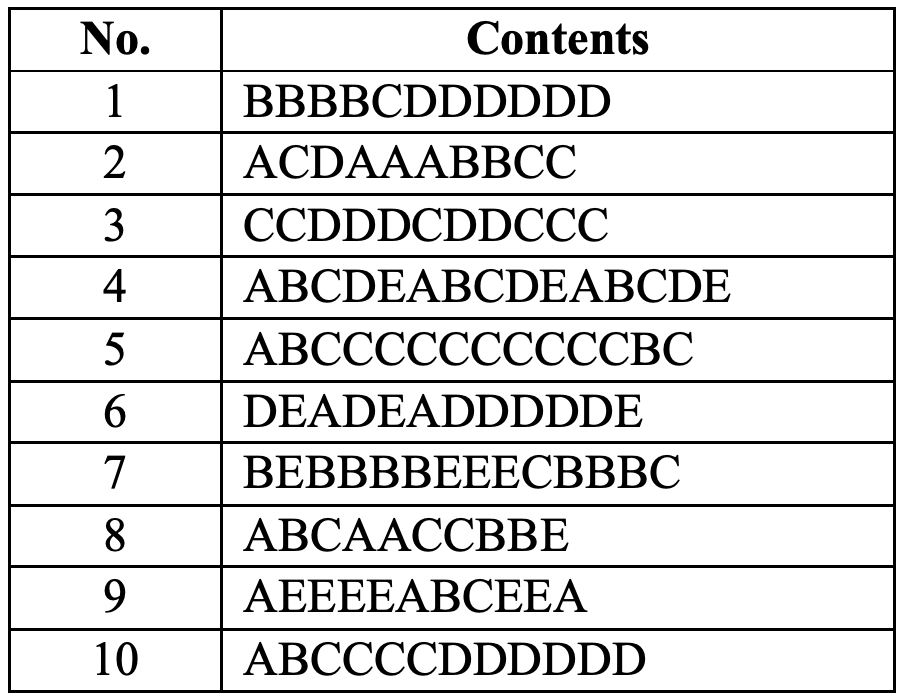
**Họ và tên: Lê Đức Huy**

**MSSV: 1810166**



**Câu 1:** Xây dựng ma trận tf.idf

**Ma trận tf:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **1** | 0 | 4 | 1 | 6 | 0 |
| **2** | 4 | 2 | 3 | 1 | 0 |
| **3** | 0 | 0 | 6 | 5 | 0 |
| **4** | 3 | 3 | 3 | 3 | 3 |
| **5** | 1 | 2 | 11 | 0 | 0 |
| **6** | 2 | 0 | 0 | 7 | 3 |
| **7** | 0 | 8 | 2 | 0 | 4 |
| **8** | 3 | 3 | 3 | 0 | 1 |
| **9** | 3 | 1 | 1 | 0 | 6 |
| **10** | 1 | 1 | 4 | 6 | 0 |

Phần tử lớn nhất trong ma trận = 11. Chuẩn hóa ma trận tf bằng cách chia mỗi phần tử cho 11.

**Ma trận tf sau khi chuẩn hóa (làm tròn 2 chữ số thập phân):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **1** | 0 | 0.36 | 0.09 | 0.55 | 0 |
| **2** | 0.36 | 0.18 | 0.27 | 0.09 | 0 |
| **3** | 0 | 0 | 0.55 | 0.45 | 0 |
| **4** | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 |
| **5** | 0.09 | 0.18 | 1.00 | 0 | 0 |
| **6** | 0.18 | 0 | 0 | 0.64 | 0.27 |
| **7** | 0 | 0.73 | 0.18 | 0 | 0.36 |
| **8** | 0.27 | 0.27 | 0.27 | 0 | 0.09 |
| **9** | 0.27 | 0.09 | 0.09 | 0 | 0.55 |
| **10** | 0.09 | 0.09 | 0.36 | 0.55 | 0 |

Công thức idf = idf = log2(N/df) = log2(10/df)

**Ma trận idf:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **idf** | log2(10/7)  = 0.51 | log2(10/8)  = 0.32 | log2(10/9)  = 0.15 | log2(10/6)  = 0.74 | log2(10/5)  = 1.00 |

**Ma trận tf.idf:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **1** | 0 | 0.12 | 0.01 | 0.41 | 0 |
| **2** | 0.18 | 0.06 | 0.04 | 0.07 | 0 |
| **3** | 0 | 0 | 0.08 | 0.33 | 0 |
| **4** | 0.14 | 0.09 | 0.04 | 0.20 | 0.27 |
| **5** | 0.05 | 0.06 | 0.15 | 0 | 0 |
| **6** | 0.09 | 0 | 0 | 0.47 | 0.27 |
| **7** | 0 | 0.23 | 0.03 | 0 | 0.36 |
| **8** | 0.14 | 0.09 | 0.04 | 0 | 0.09 |
| **9** | 0.14 | 0.03 | 0.01 | 0 | 0.55 |
| **10** | 0.05 | 0.03 | 0.05 | 0.41 | 0 |

MSSV: 181016**6**. Từ câu 2 trở đi làm dữ liệu ứng với hàng 6: DEADEADDDDDE

Câu 2: File Excel