**Intelligent Data Analysis 2020 - Exercise sheet 1**

Question-1:

Question-2:

According to the information provided in the question, we have:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Term(t)** | **IDF(t)** | **Ftd1** | **Ftd2** | **W(td1)** | **W(td2)** | **W(td1)2** | **W(td2)2** |
| action | 0.4 | 0 | 1 | 0 | 0.4 | 0 | 0.16 |
| cancel | 0.6 | 0 | 1 | 0 | 0.6 | 0 | 0.36 |
| cause | 0.3 | 0 | 1 | 0 | 0.3 | 0 | 0.09 |
| christmas | 1.5 | 1 | 1 | 1.5 | 1.5 | 2.25 | 2.25 |
| delay | 0.8 | 1 | 1 | 0.8 | 0.8 | 0.64 | 0.64 |
| england | 2.1 | 0 | 0 | 0 | 0 | 0 | 0 |
| industry | 1.6 | 0 | 1 | 0 | 1.6 | 0 | 2.56 |
| peak | 0.6 | 1 | 0 | 0.6 | 0 | 0.36 | 0 |
| period | 0.5 | 1 | 0 | 0.5 | 0 | 0.25 | 0 |
| rail | 2.2 | 1 | 0 | 2.2 | 0 | 4.84 | 0 |
| south | 0.8 | 1 | 1 | 0.8 | 0.8 | 0.64 | 0.64 |
| train | 1.9 | 1 | 2 | 1.9 | 3.8 | 3.61 | 14.44 |
|  |  |  |  |  |  | 12.59 | 21.14 |

||d1|| = 12.59

||d2|| = 21.14

1. Therefore TF-IDF Sim:

( Wchristmas.d1 . Wchristmas d2 +

Wdelay.d1 . Wdelay d2 +

Wsouth.d1 . Wsouth d2 +

Wtrain.d1 . Wtrain d2 ) / ||d1|| \* ||d2||

= (1.5 \* 1.5) + (0.8 \* 0.8) + (0.8 \* 0.8) + (1.9 \* 3.8) / (12.59 \* 21.14)

= 3.557

1. Therefore the document vectors are:
   1. Vec(d1) = Vec(d2) =