*Business Analytics with R (Prof. Andrew Lee)*

**Homework 4: Abhishek Arya**

**Net id: axa220149**

**Please use the following corpus which has 10 documents to answer the following questions.**

|  |  |
| --- | --- |
| **Document** | **Contents** |
| 1 | very well done. watch the family edition |
| 2 | great |
| 3 | other than the bad language it's not much worse than any Disney film |
| 4 | utterly disappointing & disgusting |
| 5 | watch to satisfy curiosity, not entertaining, plot very predictable, special effects fantastic |
| 6 | 9yr old daughter loved it |
| 7 | worst movie I have ever seen!!! |
| 8 | it’s an interesting movie |
| 9 | one of the best |
| 10 | worth the investment in time |

**Q1. (1 point) [short answer]** What is the value of **term frequency (tf)** for the word **‘watch’** for document 1?

**Answer) 1**

**Q2. (1 point) [short answer]** What is the value of **term frequency (tf)** for the word ‘**watch**’ for document 10?

**Answer) 0**

**Q3. (1 point) [short answer]** What is the value of **document frequency (df)** for the word ‘**watch**’?

**Answer) 2**

**Q4. (1 point) [short answer]** What is the value of **term frequency – inverse document frequency (tfidf)** for the word ‘**watch**’ for document 1?

**Answer) 0.6989**

*Business Analytics with R (Prof. Andrew Lee)*

**Q5. (1 point) [short answer]** What is the value of **term frequency – inverse document frequency (tfidf)** for the word ‘**watch**’ for document 10?

**Answer) 0**

**Q6. (1 point) [short answer]** Considering the answer to questions 4 and 5, which document (i.e., document 1 or document 10) perceives the word ‘**watch**’ as more important?

**Answer) Document 1**

**Q7. (2 point) [short answer]** Fill out the blank of the following document-term matrix with the value of **term frequency – inverse document frequency (tf-idf)** for the word ‘**watch**’.

|  |  |
| --- | --- |
| **Document** | **watch** |
| 1 | **0.6989** |
| 2 | **0** |
| 3 | **0** |
| 4 | **0** |
| 5 | **0.6989** |
| 6 | **0** |
| 7 | **0** |
| 8 | **0** |
| 9 | **0** |
| 10 | **0** |