­­­­­Timothy Mah

8750634

txamah001@mymail.sim.edu.sg

CSIT375 Artificial Intelligence for Cybersecurity **quiz2 Q3**

(a)

Tokenisation steps

1. Transform to lower case
2. Remove punctuation
3. Remove stop words
4. Correct spelling errors or abbreviations
5. Stemming

|  |  |
| --- | --- |
|  | Tokens |
| Email 1 | content, filter, model |
| Email 2 | study, spam, filter, spam, filter, subfield, content, filter |
| Email 3 | filter, spam, interest, interest |

(b)

Term frequency matrix

|  |  |  |  |
| --- | --- | --- | --- |
| tokens | Email 1 | Email 2 | Email 3 |
| content | 1 | 1 | 0 |
| filter | 1 | 3 | 1 |
| model | 1 | 0 | 0 |
| study | 0 | 1 | 0 |
| spam | 0 | 2 | 1 |
| subfield | 0 | 1 | 0 |
| interest | 0 | 0 | 2 |

(c)

Feature vectors (TF-IDF scores)

1. Apply log function to term frequency for simpler calculation. 1 + ln(value) , for all values not 0

|  |  |  |  |
| --- | --- | --- | --- |
| tokens | Email 1 | Email 2 | Email 3 |
| content | 1 | 1 | 0 |
| filter | 1 | 2.0986 | 1 |
| model | 1 | 0 | 0 |
| study | 0 | 1 | 0 |
| spam | 0 | 1.6931 | 1 |
| subfield | 0 | 1 | 0 |
| interest | 0 | 0 | 1.6931 |

1. Calculate IDF of every term
   1. IDF of one term = ln( no. of emails/ no. of emails that contain the term)

IDF scores of each term

|  |  |
| --- | --- |
| content | Ln(3/2) = 0.405 |
| filter | Ln(3/3) = 0 |
| model | Ln(3/1) = 1.099 |
| study | Ln(3/1) = 1.099 |
| spam | Ln(3/2) = 0.405 |
| subfield | Ln(3/1) = 1.099 |
| interest | Ln(3/1) = 1.099 |

1. Calculate TF-IDF score of each term
   1. TF-IDF of one term = TF x IDF

TF-IDF table

|  |  |  |  |
| --- | --- | --- | --- |
| tokens | Email 1 | Email 2 | Email 3 |
| content | 1 x 0.405 = 0.405 | 1 x 0.405 = 0.405 | 0 |
| filter | 0 | 0 | 0 |
| model | 1 x 1.099 = 1.099 | 0 | 0 |
| study | 0 | 1 x 1.099 = 1.099 | 0 |
| spam | 0 | 1.6931 x 0.405 = 0.6857 | 1 x 0.405 = 0.405 |
| subfield | 0 | 1 x 1.099 = 1.099 | 0 |
| interest | 0 | 0 | 1.6931 x 1.099 = 1.8607 |