**PS 3**

**Class ID: 20**

**Name: Rakesh Reddy Pallepati**

1.Ans :

After processing each and every individual document by stopwords functions Tokenization the output will be as follows:

Doc:1

The researchers focus computational phenotyping produce disease prediction models machine learning statistical tools

Doc:2

The researchers develop tools Bayesian statistical information generate causal models large complex phenotyping datasets

Doc:3

The researchers build computational information engine uses machine learning combine gene function gene interaction information disparate genomic data sources

Multiword Terms: These terms are extracted using the n-gram approach where we need to give the value as 2 or 3 .so that it can produce the output with two words from that we can extract the most useful words related to context.

computational phenotyping,

disease prediction ,

machine learning,

statistical tools,

Bayseian statistical information,

causal models,

phenotyping datasets,

computational information engine,

genomic data sources

gene function

gene interaction

2.Ans :

TF(t) = (N0.of times term t appears in a document) / (Total no. of terms in the document).

IDF(t) = log(1+(Total number of documents / Number of documents with term t in it)).

Doc1#:

Order word count Occurrences TF(%) IDF

1. produce 1 7.6923 0.6020

2. learning 1 7.6923 0.3979

3. researchers 1 7.6923 0.3979

4. models 1 7.6923 0.3979

5. computational 1 7.6923 0.3979

6. tools 1 7.6923 0.3979

7. phenotyping 1 7.6923 0.3979

8. statistical 1 7.6923 0.3979

9. prediction 1 7.6923 0.6020

10. focus 1 7.6923 0.6020

11. machine 1 7.6923 0.3979

12. disease 1 7.6923 0.6020

13. the 1 7.6923 0.3010

Doc2#:

Order Word Occurrences TF(%) IDF

1. information 1 7.1429 0.3979

2. causal 1 7.1429 0.3979

3. generate 1 7.1429 0.3979

4. bayesian 1 7.1429 0.6020

5. researchers 1 7.1429 0.3979

6. develop 1 7.1429 0.6020

7. models 1 7.1429 0.6020

8. datasets 1 7.1429 0.3979

9. tools 1 7.1429 0.6020

10. phenotyping 1 7.1429 0.3979

11. statistical 1 7.1429 0.3979

12. complex 1 7.1429 0.6020

13. large 1 7.1429 0.6020

14. the 1 7.1429 0.3010

Doc3#:

Order word Occurrences TF(%) IDF

1. information 2 10.5263 0.3979

2. gene 2 10.5263 0.6020

3. learning 1 5.2632 0.3979

4. engine 1 5.2632 0.6020

5. disparate 1 5.2632 0.6020

6. genomic 1 5.2632 0.6020

7. researchers 1 5.2632 0.3979

8. interaction 1 5.2632 0.6020

9. computational 1 5.2632 0.6020

10. combine 1 5.2632 0.6020

11. uses 1 5.2632 0.6020

12. data 1 5.2632 0.6020

13. sources 1 5.2632 0.6020

14. function 1 5.2632 0.6020

15. machine 1 5.2632 0.3979

16. the 1 5.2632 0.3010

17. build 1 5.2632 0.6020

**Inverse Document Frequency:**

Total number of documents = 3

IDF for the words are:

‘the’ – log\_e(3/3) = 0

‘researchers’ – log\_e(3/3) = 0

‘focus’ – log\_e(3/1) = 1.09

‘computational’ – log\_e(3/2) = 0.40

‘phenotyping’ – log\_e(3/2) = 0.40

‘produce’ – log\_e(3/1) = 1.09

‘disease’ – log\_e(3/1) = 1.09

‘prediction’ – log\_e(3/3) = 0

‘models’ – log\_e(3/1) = 1.09

‘machine’ – log\_e(3/2) = 0.40 ‘learning’ – log\_e(3/2) = 0.40

‘statistical’ – log\_e(3/2) = 0.40

‘tools’ – log\_e(3/1) = 1.09

‘develop’ – log\_e(3/1) = 1.09

‘Bayesian’ – log\_e(3/1) = 1.09

‘information’ – log\_e(3/2) = 0.40

‘generate’ – log\_e(3/1) = 1.09

‘causal’ – log\_e(3/1) = 1.09

‘large’ – log\_e(3/1) = 1.09

‘complex’ – log\_e(3/1) = 1.09

‘datasets’ – log\_e(3/1) = 1.09

‘build’ – log\_e(3/1) = 1.09

‘engine’ – log\_e(3/1) = 1.09

‘uses’ – log\_e(3/1) = 1.09

‘combine’ – log\_e(3/1) = 1.09 ‘gene’ – log\_e(3/2) = 0.40

‘function’ – log\_e(3/1) = 1.09

‘interaction’ – log\_e(3/1) = 1.09

‘disparate’ – log\_e(3/1) = 1.09

‘genomic’ – log\_e(3/1) = 1.09

‘data’ – log\_e(3/1) = 1.09

‘sources’ – log\_e(3/1) = 1.09

Term Weights:

Term Weight for ‘the’ – 0

Term Weight for ‘researchers’ – 0

Term Weight for ‘focus’ – 0.0769 \* 1.09 = 0.083

Term Weight for ‘computational’ – 0.0769 \* 0.40 = 0.030

Term Weight for ‘phenotyping’ – 0.0769 \* 0.40 = 0.030

Term Weight for ‘produce’ – 0.0769 \* 1.09 = 0.083

Term Weight for ‘disease’ 0.0769 \* 1.09 = 0.083

Term Weight for ‘prediction’ – 0.0769 \* 0 = 0

Term Weight for ‘models’ – 0.0769 \* 1.09 = 0.083

Term Weight for ‘machine’ – 0.0769 \* 0.40 = 0.030

Term Weight for ‘learning’ – 0.0769 \* 0.40 = 0.030

Term Weight for ‘statistical’ – 0.0769 \* 0.40 = 0.030

Term Weight for ‘tools’ - 0.0769 \* 1.09 = 0.083

Term Weight for ‘develop’ – 0.07142 \* 1.09 = 0.077

Term Weight for ‘Bayesian’ – 0.07142 \* 1.09 = 0.077

Term Weight for ‘information’ – 0.07142 \* 0.40 = 0.028

Term Weight for ‘generate’ – 0.07142 \* 1.09 = 0.077

Term Weight for ‘causal’ –0.07142 \* 1.09 = 0.077

Term Weight for ‘large’ – 0.07142 \* 1.09 = 0.077

Term Weight for ‘complex’ – 0.07142 \* 1.09 = 0.077

Term Weight for ‘datasets’ – 0.07142 \* 1.09 = 0.077

Term Weight for ‘build’ – 0.05263 \* 1.09 = 0.057

Term Weight for ‘engine’ – 0.05263 \* 1.09 = 0.057

Term Weight for ‘uses’ – 0.05263 \* 1.09 = 0.057

Term Weight for ‘combine’ – 0.05263 \* 1.09 = 0.057

Term Weight for ‘gene’ – 0.05263 \* 0.40 = 0.021

Term Weight for ‘function’ – 0.05263 \* 1.09 = 0.057

Term Weight for ‘interaction’ – 0.05263 \* 1.09 = 0.057

Term Weight for ‘disparate’ – 0.05263 \* 1.09 = 0.057

Term Weight for ‘genomic’ – 0.05263 \* 1.09 = 0.057

Term Weight for ‘data’ – 0.05263 \* 1.09 = 0.057

Term Weight for ‘sources’ – 0.05263 \* 1.09 = 0.057

Document Vector:

|  |  |  |  |
| --- | --- | --- | --- |
| Term | Doc1 | Doc2 | Doc3 |
| the | 1 | 1 | 1 |
| Researchers | 1 | 1 | 1 |
| Focus | 1 | 0 | 0 |
| Computational | 1 | 0 | 1 |
| Phenotyping | 1 | 1 | 0 |
| Produce | 1 | 0 | 0 |
| Disease | 1 | 0 | 0 |
| Models | 1 | 0 | 0 |
| Machine | 1 | 0 | 1 |
| Learning | 1 | 0 | 1 |
| Statistical | 1 | 1 | 0 |
| Tools | 1 | 0 | 0 |
| Develop | 0 | 1 | 0 |
| Bayesian | 0 | 1 | 0 |
| Information | 0 | 1 | 0 |
| Generate | 0 | 1 | 0 |
| Causal | 0 | 1 | 0 |
| Large | 0 | 1 | 0 |
| Complex | 0 | 1 | 0 |
| Datasets | 0 | 1 | 0 |
| Build | 0 | 0 | 1 |
| Engine | 0 | 0 | 1 |
| uses | 0 | 0 | 1 |
| Combine | 0 | 0 | 1 |
| Gene | 0 | 0 | 1 |
| Function | 0 | 0 | 1 |
| Interaction | 0 | 0 | 1 |
| Disparate | 0 | 0 | 1 |
| genomic | 0 | 0 | 1 |
| Data | 0 | 0 | 1 |
| sources | 0 | 0 | 1 |