**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