Problem set-3

Name: Syed Moin
class id: 28

1A) Removal of stopwords and punctuation:

o/p:

DOC 1:
Reasearchers
focus
computational
phenotyping
produce
disease
prediction
models
machine
learning
statistical
tools.

DOC 2:
Researchers
develop
tools
Bayesian
statistical
information
generate
causalmodels
large
complex
phenotyping
datasets.

DOC 3:
researchers
build
computational
information
engine
uses
machine
learning
combine
gene
function
gene
integration
information
disparate
genomic
data
sources.

N-gram after removal of stop words: N=3

Doc 1:

researchers focus computational

focus computational phenotyping

computational phenotyping produce

phenotyping produce disease

produce disease prediction

disease prediction models

prediction models machine

models machine learning

machine learning statistical

learning statistical tools

Doc 2:

reasearchers develop tools

develop tools bayesian

tools Bayesian statistical

Bayesian statistical information

statistical information generate

information generate casual

generate casual models.

casual models large

models large complex

large complex phenotyping

complex phenotyping datasets.

DOC3:

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

(B) Term frequency:

$$TF(t) = \frac{\text{No. of times term } t \text{ appears in a document}}{\text{Total no of terms in the document.}}$$

Inverse document frequency:

$$IDF(t) = \log_e \left(\frac{\text{Total no. of documents}}{\text{No of documents with term } t \text{ in it}} \right)$$

	D1	D2	D3
researchers	1	1	1
field	1	0	0
computational	1	0	1
phenotyping	1	1	0
produce	1	0	0
disease	1	0	0
prediction	1	0	0
models	1	1	0
machine	1	0	1
learning	1	0	1
statistical	1	1	0
tools	1	1	0
develop	0	1	0
Bayesian	0	1	0
information	0	1	2
generate	0	1	0
causal	0	1	0
large	0	1	0
complex	0	1	0
datasets	0	1	0
build	0	0	1
sources	0	0	1
data	0	0	1
gene	0	0	2
genomic	0	0	1

TF-IDF values for each terms in Documents 1:

1) researches ! $TF = \frac{1}{12}$ $IDF = \log_e\left(\frac{3}{2}\right)$

$TF \cdot IDF = \frac{1}{12} \times \log_e\left(\frac{3}{2}\right) = 0.0146$

2) Focus : $TF = \frac{1}{12}$, $IDF = \log_e\left(\frac{3}{1}\right)$; $TF \cdot IDF = 0.039$

3) computational ; $TF = \frac{1}{12}$, $IDF = \log_e\left(\frac{3}{2}\right)$; $TF\text{-}IDF = 0.0146$

4) Phenotyping ·

4) For produce , disease , prediction

$TF = \frac{1}{12}$, $IDF = \log_e\left(\frac{3}{1}\right) = 3$; $TF\text{-}IDF = 0.039$

5) For models, machine , learning , statistical , tools

$TF = \frac{1}{12}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = 0.0146$

6) For remaining all terms which are not present in D1

$TF = 0$, then $TF \cdot IDF = 0$.

For Document 2:

1) Develop, Bayesian

$TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{1}\right) = 0.477$, $TF\text{-}IDF = 0.036$

2) Information , tools , statistical

$TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = 0.0135$

3) generate , large , datasets , causal, complex

$TF = \frac{1}{13}$, $IDF = \log(\frac{3}{1})$ $TF-IDF = 0.036$

4) phenotyping , models

$TF = \frac{1}{13}$, $IDF = \log(\frac{3}{2})$, $TF-IDF = 0.0135$

Doument 3 :

1) researchers.

$TF = \frac{1}{18}$ $IDF = \log(\frac{3}{3})$, $TF-IDF = 0$

2) build

$TF = \frac{1}{18}$ $IDF = \log(\frac{3}{1})$, $TF-IDF = 0.026$

3) For engine , uses , machine, combine , function , interaction, data , sources , genomic , disparate

$TF = \frac{1}{18}$. $IDF = \log(\frac{3}{1})$ $TF-IDF = 0.026$

4) Gene

$TF = \frac{2}{18}$, $IDF = \log(\frac{3}{1})$ $TF-IDF = 0.0529$

5) learning

$TF = \frac{1}{18}$ $IDF = \log(\frac{3}{2})$ $TF-IDF = 0.978.$