Problem Set 3

Name: SANDEEP PABOLU
Class Id: 19

Question 1:

a) First remove stop words and punctuation; detect manually multi-word terms (using NGram or POS Tagging/Chunking); parse manually the documents and select the terms from the given 3 documents and created the dictionary (list of terms).

Ans:

1. The given statement is:

"The researchers **will** focus **on** computational phenotyping **and will** produce disease prediction models **from** machine learning **and** statistical tools."

On removal of stop words: will, on,and,will,from,and

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

2. The given statement is:

"The researchers **will** develop tools **that use** Bayesian statistical information **to** generate causal models **from** large **and** complex phenotyping datasets."

On removal of stop words:

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

3. The given statement is:

"The researchers **will** build **a** computational information engine **that** uses machine learning **to** combine gene function **and** gene interaction information **from** disparate genomic data sources."

On removal of stop words:

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

Multi-Word terms in all documents are:

The	3
Researchers	3
Information	3
machine	2
Gene	2
tools	2
Statistical	2
learning	2
Models	2
Phenotyping	2
computational	2

Dictionary D = {the, researchers, information, machine, gene, tools, statistical, learning, models, phenotyping, computational}

Question 2:

b) Create the document vectors by computing TF-IDF weights. Show how to compute the TF-IDF weights for terms. For each form of weighting list the document vectors in the following format:

Answer:

For 1:

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

Total number of terms in the document = 13

The term frequencies for:
'the' – 1/13 = 0.0769
'researchers' – 1/13 = 0.0769
'focus' – 1/13 = 0.0769
'computational' – 1/13 = 0.0769
'phenotyping' – 1/13 = 0.0769
'produce' – 1/13 = 0.0769
'disease' – 1/13 = 0.0769
'prediction' – 1/13 = 0.0769
'models' – 1/13 = 0.0769
'machine' – 1/13 = 0.0769
'learning' – 1/13 = 0.0769

‘statistical’ – 1/13 = 0.0769
‘tools’ – 1/13 = 0.0769

For 2:
The researchers develop tools Bayesian statistical information generate causal models large complex phenotyping datasets

Total number of terms in the document = 14
The Term frequencies for:
‘the’ – 1/14 = 0.07142
‘researchers’ – 1/14 = 0.07142
‘develop’ – 1/14 = 0.07142
‘tools’ – 1/14 = 0.07142
‘Bayesian’ – 1/14 = 0.07142
‘statistical’ – 1/14 = 0.07142
‘information’ – 1/14 = 0.07142
‘generate’ – 1/14 = 0.07142
‘causal’ – 1/14 = 0.07142
‘models’ – 1/14 = 0.07142
‘large’ – 1/14 = 0.07142
‘complex’ – 1/14 = 0.07142
‘phenotyping’ – 1/14 = 0.07142
‘datasets’ – 1/14 = 0.07142

For 3:
The researchers build computational information engine uses machine learning combine gene function gene interaction information disparate genomic data sources

Total number of terms in the document = 19

The Term frequencies for:
‘the’ – 1/19 = 0.05263
‘researchers’ – 1/19 = 0.05263
‘build’ – 1/19 = 0.05263
‘computational’ – 1/19 = 0.05263
‘information’ – 2/13 = 0.10526
‘engine’ – 1/19 = 0.05263
‘uses’ – 1/19 = 0.05263
‘machine’ – 1/19 = 0.05263
‘learning’ – 1/19 = 0.05263
‘combine’ – 1/19 = 0.05263
‘gene’ – 2/13 = 0.10526
‘function’ – 1/19 = 0.05263

‘interaction’ – 1/19 = 0.05263
‘disparate’ – 1/19 = 0.05263
‘genomic’ – 1/19 = 0.05263
‘data’ – 1/19 = 0.05263
‘sources’ – 1/19 = 0.05263

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