Problem Set 3

Name: Sai Sriharsha Sudulaguntla

Class Id: 25

Information Retrieval (Text Mining) with TF-IDF

**Doc #1:**

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

**Doc #2:**

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

**Doc #3:**

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.

**Question 1:**

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

Doc #1: The researchers will focus on computational phenotyping and will produce disease prediction models from machine learning and statistical tools.

On removing stop words and punctuation: The researchers focus computational phenotyping produce disease prediction models machine learning statistical tools

Doc #2: The researchers will develop tools that use Bayesian statistical information to generate causal models from large and complex phenotyping datasets.

On removing stop words and punctuation:The researchers develop tools Bayesian statistical information generate causal models large complex phenotyping datasets

Doc #3: 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 removing stop words and punctuation: The researchers build computational information engine uses machine learning combine gene function gene interaction information disparate genomic data sources.

**Multi-Word terms in three documents combined:**

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:**

1. 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:

Doc #1:

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

TF(Term Frequency) – Doc1:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

Total number of terms in the document = 13

Term frequency for **‘the’** – 1/13 = 0.0769

Term frequency for **‘researchers’** – 1/13 = 0.0769

Term frequency for **‘focus’** – 1/13 = 0.0769

Term frequency for **‘computational’** – 1/13 = 0.0769

Term frequency for **‘phenotyping’** – 1/13 = 0.0769

Term frequency for **‘produce’** – 1/13 = 0.0769

Term frequency for **‘disease’** – 1/13 = 0.0769

Term frequency for **‘prediction’** – 1/13 = 0.0769

Term frequency for **‘models’** – 1/13 = 0.0769

Term frequency for **‘machine’** – 1/13 = 0.0769

Term frequency for **‘learning’** – 1/13 = 0.0769

Term frequency for **‘statistical’** – 1/13 = 0.0769

Term frequency for **‘tools’** – 1/13 = 0.0769

**Doc #2:**

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

TF(Term Frequency) – Doc2:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

Total number of terms in the document = 14

Term frequency for **‘the’** – 1/14 = 0.07142

Term frequency for **‘researchers’** – 1/14 = 0.07142

Term frequency for **‘develop’** – 1/14 = 0.07142

Term frequency for **‘tools’** – 1/14 = 0.07142

Term frequency for **‘Bayesian’** – 1/14 = 0.07142

Term frequency for **‘statistical’** – 1/14 = 0.07142

Term frequency for **‘information’** – 1/14 = 0.07142

Term frequency for **‘generate’** – 1/14 = 0.07142

Term frequency for **‘causal’** – 1/14 = 0.07142

Term frequency for **‘models’** – 1/14 = 0.07142

Term frequency for **‘large’** – 1/14 = 0.07142

Term frequency for **‘complex’** – 1/14 = 0.07142

Term frequency for **‘phenotyping’** – 1/14 = 0.07142

Term frequency for **‘datasets’** – 1/14 = 0.07142

**Doc #3:**

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

TF(Term Frequency) – Doc3:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

Total number of terms in the document = 19

Term frequency for **‘the’** – 1/19 = 0.05263

Term frequency for **‘researchers’** – 1/19 = 0.05263

Term frequency for **‘build’** – 1/19 = 0.05263

Term frequency for **‘computational’** – 1/19 = 0.05263

Term frequency for **‘information’** – 2/13 = 0.10526

Term frequency for **‘engine’** – 1/19 = 0.05263

Term frequency for **‘uses’** – 1/19 = 0.05263

Term frequency for **‘machine’** – 1/19 = 0.05263

Term frequency for **‘learning’** – 1/19 = 0.05263

Term frequency for **‘combine’** – 1/19 = 0.05263

Term frequency for **‘gene’** – 2/13 = 0.10526

Term frequency for **‘function’** – 1/19 = 0.05263

Term frequency for **‘interaction’** – 1/19 = 0.05263

Term frequency for **‘disparate’** – 1/19 = 0.05263

Term frequency for **‘genomic’** – 1/19 = 0.05263

Term frequency for **‘data’** – 1/19 = 0.05263

Term frequency for **‘sources’** – 1/19 = 0.05263

Inverse Document Frequency:

Total number of documents = 3

IDF(t) = log\_e(Total number of documents / Number of documents with term t in it)

IDF for **‘the’** – log\_e(3/3) = 0

IDF for **‘researchers’** – log\_e(3/3) = 0

IDF for **‘focus’** – log\_e(3/1) = 1.09

IDF for **‘computational’** – log\_e(3/2) = 0.40

IDF for **‘phenotyping’** – log\_e(3/2) = 0.40

IDF for **‘produce’** – log\_e(3/1) = 1.09

IDF for **‘disease’** – log\_e(3/1) = 1.09

IDF for **‘prediction’** – log\_e(3/3) = 0

IDF for **‘models’** – log\_e(3/1) = 1.09

IDF for **‘machine’** – log\_e(3/2) = 0.40

IDF for **‘learning’** – log\_e(3/2) = 0.40

IDF for **‘statistical’** – log\_e(3/2) = 0.40

IDF for **‘tools’** – log\_e(3/1) = 1.09

IDF for **‘develop’** – log\_e(3/1) = 1.09

IDF for **‘Bayesian’** – log\_e(3/1) = 1.09

IDF for **‘information’** – log\_e(3/2) = 0.40

IDF for **‘generate’** – log\_e(3/1) = 1.09

IDF for **‘causal’** – log\_e(3/1) = 1.09

IDF for **‘large’** – log\_e(3/1) = 1.09

IDF for **‘complex’** – log\_e(3/1) = 1.09

IDF for **‘datasets’** – log\_e(3/1) = 1.09

IDF for **‘build’** – log\_e(3/1) = 1.09

IDF for **‘engine’** – log\_e(3/1) = 1.09

IDF for **‘uses’** – log\_e(3/1) = 1.09

IDF for **‘combine’** – log\_e(3/1) = 1.09

IDF for **‘gene’** – log\_e(3/2) = 0.40

IDF for **‘function’** – log\_e(3/1) = 1.09

IDF for **‘interaction’** – log\_e(3/1) = 1.09

IDF for **‘disparate’** – log\_e(3/1) = 1.09

IDF for **‘genomic’** – log\_e(3/1) = 1.09

IDF for **‘data’** – log\_e(3/1) = 1.09

IDF for **‘sources’** – log\_e(3/1) = 1.09

Term Weights:

Term Weight = TF \* IDF

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 = 0.05**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 |
| Prediction | 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 |